Other Minds

HOW HUMANS BRIDGE THE DIVIDE BETWEEN SELF AND OTHERS

edited by Bertram F. Malle & Sara D. Hodges

OTHER MINDS

This page intentionally left blank

Other Minds

How Humans Bridge the Divide between Self and Others

Edited by

BERTRAM F. MALLE
SARA D. HODGES

THE GUILFORD PRESS
New York London

© 2005 The Guilford Press
A Division of Guilford Publications, Inc.
72 Spring Street, New York, NY 10012
www.guilford.com

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Other minds : how humans bridge the divide between self and others / edited by
Bertram F. Malle, Sara D. Hodges.
p. cm.
Includes bibliographical references and index.
ISBN 1-59385-187-1 (hardcover)
1. Social perception. 2. Other minds (Theory of knowledge) I. Malle, Bertram F.
II. Hodges, Sara D.
BF323.S63O89 2005
121′.2—dc22
2005009644

About the Editors

Bertram F. Malle, PhD, is Associate Professor of Psychology at the University of Oregon and the recipient of a Society of Experimental Social Psychology Dissertation Award and a National Science Foundation Career Award. His research examines the cognitive tools that humans bring to social interaction, such as the folk concept of intentionality, inferences of mental states, and explanations of behavior. Dr. Malle is coeditor of two other volumes, *Intentions and Intentionality* 2001, MIT Press) and *The Evolution of Language Out of Pre-Language* (2002, Benjamins), and the author of *How the Mind Explains Behavior* (2004, MIT Press).

Sara D. Hodges, PhD, is Associate Professor of Psychology at the University of Oregon. Her research, which has been published in scholarly journals and edited volumes, explores the role of the self in people's perceptions of others, with a particular emphasis on empathy, projection, and social comparison. Dr. Hodges also studies how people construct evaluations and preferences in social contexts. Her research has received funding from the National Science Foundation, and she has been the recipient of two Rippey Innovative Teaching Awards at the University of Oregon.

This page intentionally left blank

Contributors

Daniel R. Ames, PhD, Columbia Business School, Columbia University, New York, New York

Janet Wilde Astington, PhD, Ontario Institute for Studies in Education, University of Toronto, Toronto, Ontario, Canada

Jodie A. Baird, PhD, Department of Psychology, Villanova University, Villanova, Pennsylvania

Marjorie Barker, PhD, Department of Linguistics, University of Oregon, Eugene, Oregon

Dale J. Barr, PhD, Department of Psychology, University of California, Riverside, California

Radu J. Bogdan, PhD, Department of Philosophy, Tulane University, New Orleans, Louisiana

Mark H. Davis, PhD, Department of Psychology, Eckerd College, St. Petersburg, Florida

Jean Decety, PhD, Institute for Learning and Brain Sciences, University of Washington, Seattle, Washington

Diego Fernandez-Duque, PhD, Department of Psychology, Villanova University, Villanova, Pennsylvania

Eva Filippova, MA, Ontario Institute for Studies in Education, University of Toronto, Toronto, Ontario, Canada

Susan R. Fussell, PhD, Human–Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania

Darren Gergle, MS, Human–Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania

T. Givón, PhD, Department of Linguistics, University of Oregon, Eugene, Oregon

Sara D. Hodges, PhD, Department of Psychology and Institute of Cognitive and Decision Sciences, University of Oregon, Eugene, Oregon

Daniel D. Hutto, DPhil, School of Humanities, University of Hertfordshire, Herts, United Kingdom

William Ickes, PhD, Department of Psychology, University of Texas, Arlington, Texas

Boaz Keysar, PhD, Department of Psychology, University of Chicago, Chicago, Illinois

Robert E. Kraut, PhD, Human–Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania

Anton Kühberger, PhD, Department of Psychology, University of Salzburg, Salzburg, Austria

Robyn Langdon, PhD, Macquarie Center for Cognitive Science, Macquarie University, Sydney, Australia

George Loewenstein, PhD, Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania

Bertram F. Malle, PhD, Department of Psychology and Institute of Cognitive and Decision Sciences, University of Oregon, Eugene, Oregon

Lynn C. Miller, PhD, Department of Psychology, University of Southern California, Los Angeles, California

Louis J. Moses, PhD, Department of Psychology, University of Oregon, Eugene, Oregon

Minda Oriña, PhD, Department of Communication, Michigan State University, East Lansing, Michigan

Josef Perner, PhD, Department of Psychology, University of Salzburg, Salzburg, Austria

Stephen J. Read, PhD, Department of Psychology, University of Southern California, Los Angeles, California

Glenn D. Reeder, PhD, Department of Psychology, Illinois State University, Normal, Illinois

Michael F. Schober, PhD, Department of Psychology, New School University, New York, New York

Leslie D. Setlock, MA, Human–Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania

Jeffry A. Simpson, PhD, Department of Psychology, University of Minnesota, Minneapolis, Minnesota

David Trafimow, PhD, Department of Psychology, New Mexico State University, Las Cruces, New Mexico

James S. Uleman, PhD, Department of Psychology, New York University, New York, New York

Leaf Van Boven, PhD, Department of Psychology, University of Colorado, Boulder, Colorado

Preface

This book not only is about "other minds"; it is a collaborative effort that drew together many other minds besides our own. The book grew out of a conference held in September 2003 at the University of Oregon, for which we sought to bring together scholars from such diverse disciplines as psychology, linguistics, decision science, philosophy, primatology, and neuroscience. The conference contributions highlighted the diversity of conceptual and methodological approaches to the multiply labeled topic of other minds, mindreading, or theory of mind, and we hope that the current volume is a good representation of this diversity.

Mindful of the difficulties readers might encounter with a project involving a variety of minds talking about other minds, we strove to keep the contributions to this book accessible, encouraging authors to think about their chapters as essays, not journal articles. Chapter authors also helped in this endeavor, as many of them anonymously commented on another author's chapter, adding greatly to our own editorial efforts.

Projects like this require not only other minds but also other sources of support. We are grateful to the University of Oregon's Institute of Cognitive and Decision Sciences and the Department of Psychology. In addition, we would like to thank some particular minds (and bodies) for helping to make the initial conference and the later volume successful: Vonda Evans, Marjorie Taylor, Lara London, Lael Rogan, Mike Myers, Christina Gamache, Nicole Selinger, Obsidians, Inc., the San Francisco Museum of Modern Art, Seymour Weingarten of The Guilford Press, and the students in the 2004 Other Minds course at the University of Oregon's Clark Honors College.

Contents

Introduction 1

PART I. QUESTIONS ABOUT THE PHENOMENON

1 Executive Functioning and Children's Theories of Mind 11
Louis J. Moses

2 Three Puzzles of Mindreading 26
Bertram F. Malle

3 A "Constituent" Approach to the Study of Perspective Taking: What Are Its Fundamental Elements? 44
Mark H. Davis

4 Starting without Theory: Confronting the Paradox of Conceptual Development 56
Daniel D. Hutto

PART II. READING BEHAVIOR, READING MINDS

5 Is There a "Social Brain"?: Lessons from Eye-Gaze Following, Joint Attention, and Autism 75
Diego Fernandez-Duque and Jodie A. Baird

6 Visual Cues as Evidence of Others' Minds in Collaborative Physical Tasks 91
Susan R. Fussell, Robert E. Kraut, Darren Gergle, and Leslie D. Setlock

7 Attributing Motives to Other People 106
Glenn D. Reeder and David Trafimow

8 Explanatory Coherence and Goal-Based Knowledge Structures in Making Dispositional Inferences 124
Stephen J. Read and Lynn C. Miller

PART III. READING ONE'S OWN MIND, READING OTHER MINDS

9 Perspective Taking as the Royal Avenue to Empathy 143
Jean Decety

10 Everyday Solutions to the Problem of Other Minds: Which Tools Are Used When? 158
Daniel R. Ames

11 Mental Simulation: Royal Road to Other Minds? 174
Josef Perner and Anton Kühberger

12 Why Self-Ascriptions Are Difficult and Develop Late 190
Radu J. Bogdan

PART IV. LANGUAGE AND OTHER MINDS

13 Language as the Route into Other Minds 209
Janet Wilde Astington and Eva Filippova

14 Representation of the Interlocutor's Mind during Conversation 223
Marjorie Barker and T. Givón

15 Conceptual Alignment in Conversation 239
Michael F. Schober

16 On the Inherent Ambiguity of Traits and Other Mental Concepts 253
James S. Uleman

PART V. LIMITS OF MINDREADING

17 Mindreading in an Exotic Case: The Normal Adult Human 271
Dale J. Barr and Boaz Keysar

18 Empathy Gaps in Emotional Perspective Taking 284
Leaf Van Boven and George Loewenstein

19 Is How Much You Understand Me in Your Head or Mine? 298
Sara D. Hodges

20 Empathic Accuracy and Inaccuracy in Close Relationships 310
William Ickes, Jeffry A. Simpson, and Minda Oriña

21 Theory of Mind in Schizophrenia 333
Robyn Langdon

Index 343

This page intentionally left blank

Introduction

The classic philosophical discussion of other minds is framed as the "other minds *problem*." The problem is whether we can ever know what is going on in another person's mind. Philosophical debates rarely provide answers but challenge us with further questions: What do we actually mean by *know,* or by *mind*, or by *ever?* None of these philosophical questions will be answered in the present book. Nonetheless, the "problem" frame is visible throughout the current attempts of empirical science to illuminate the topic of other minds. The problem ordinary people face is less one of *whether* they can know other minds but one of *how* to find out about other minds—what tools to use, what information to seek out, what conclusions to draw. The fact that there is a "problem" at all with other minds suggests that people care about what others are thinking and feeling and care about being reasonably accurate about those mental states. Of course, it's no wonder that people care—humans are social creatures, and coordinating their actions with others should be greatly facilitated by knowing their thoughts, feelings, and intentions. The adaptive value of these skills has apparently promoted a whole host of cognitive mechanisms and neural structures that all help the perceiver with the general task of grasping other minds. These adaptations in turn create the *scientist's* other minds problem: that of identifying those mechanisms and structures and delineating how they cooperate in connecting one mind to another.

Framing the question of knowing other minds as a problem also suggests that some of the attempts to grasp other minds will fail—and there is arguably ample evidence of humans' lack of perfection at this task. Who hasn't been told "You have no idea what it feels like" or bemoaned in others a resistance to adopt our perspective? And while others' thoughts seem cloudy, perplexing, and hidden, our own thoughts unceasingly proclaim themselves. Ironically, this very inescapability of our own perspective (which sometimes threatens to drown out other perspectives) can at times be one of our most useful tools in comprehending other minds.

This book joins a growing number of collaborative works that examine the fascinating capacity to represent and reason about minds. Some of them have a specific topic focus, such as action perception, intentionality, imitation, or self-control (e.g., Frith & Wolpert, 2004; Malle, Moses, & Baldwin, 2001; Meltzoff & Prinz, 2002; Zelazo, Astington, & Olson, 1999); others have examined more broadly the capacity to read minds (Byrne & Whiten, 1986; Carruthers & Smith, 1996; Davies & Stone, 1995; Frye & Moore, 1991). The present volume offers some of the latest developments in this multidisciplinary field, but it also tries to make mindreading itself the specific topic of analysis. What is it that we call mindreading, mental state inference, or mentalizing? What specific processes, capacities, and neural substrates do we subsume under these labels? How are these specific elements related, and how do they in turn relate to language, self-awareness, and social interaction?

The empirical study of the other minds problem is open to investigation in a number of disciplines. Consequently, the present collection of essays covers multiple theoretical, methodological, and disciplinary perspectives. Even so, the multiple perspectives converge on a number of important insights. Perhaps the most important one concerns the scope of the empirical other minds problem. Previous research and debates centered on the label of "theory of mind"—the cognitive system that allows an organism to grasp and reason about minds. Initially, the dominant perspective portrayed theory of mind literally as a theory, a set of rules and principles, or as a module, a set of dedicated cognitive mechanisms. Increasingly over the past decade, a competing perspective has emerged under the label of "simulation theory," which describes how one mind simulates what might be going on in the other mind and delineates a powerful process by which humans may represent other minds without necessarily relying on a module, rules, or principles.

Despite the enormous productivity of research on these two perspectives, the focus on the debate between them has constrained the study of other minds in a variety of ways. Attempts by one perspective to best the other in a winner-takes-all competition all but ignored the possibility that both strategies might be used or that other strategies might also be in play. Furthermore, a very large portion of the classic work examines inferences of beliefs and desires, keeping at a distance several related phenomena that might be subsumed under a broader conceptualization of theory of mind. For example, eye-gaze following and emotional contagion are considered precursors of a theory of mind, and, primitive as they may be, they contribute to our perceptions of and transactions with other minds. In addition, the interpretation of communicative actions, such as comprehending other people's words in light of their particular context, could be considered an important application of theory of mind.

In the present book, we have tried to broaden the scope of what previously fell under the label "theory of mind," a goal that reflects current fresh approaches being taken by researchers. Along with some of the contributing authors, we have adopted the label *mindreading* to escape the tethers of the theory-versus-simulation debate, with the advantage of labeling the activity that people perform without making specific assumptions about what processes and mechanisms underlie this activity. But it's not just labeling that is beginning to change. More and more writers recognize the whole range of phenomena that help people solve the other minds problem—from eye-gaze monitoring and automatic perception–action links at one end to interpretation of communicative meaning and full-fledged perspective taking at the other end. All these tools can be subsumed under one broader function—that of helping the organism connect and coordinate with other mindful organisms. At the same time, that doesn't mean the tools are so tightly linked that we could speak of one mechanism, module, or system. Neuroscientific evidence increasingly suggests that a number of brain systems are involved in the organism's engagement with other minds; distinct developmental paths are being discovered about distinct elements in the mindreading "tool box"; and evolutionary models may have to be differentiated to account for the distinct origin and history of some of these tools.

This volume thus offers an array of answers to a question that has been slightly altered from its philosophical roots: *How do ordinary people understand other minds?* As part of this "how" problem, the chapters in this book document the variety of cognitive and social tools that have evolved over our biological and cultural history to bridge the seeming distance to other people's minds. Some chapters specifically explore the developmental context in which these tools emerge: when and in conjunction with what other skills they become operative. Others search for the neural correlates of these tools and relate them to the critical phenomena of language and self-awareness. Some chapters describe the social functions and consequences of mindreading, especially in conversation and relationships. Others explore the limitations of human mindreaders, both in everyday contexts and in cases of psychopathology.

Part I of this book starts out with some of the guiding questions in the empirical study of the other minds problem. After two decades of intensive research, after filling in dots on a large map, we now have to examine what the contours of the map are, where its boundaries lie, and what this map is ultimately a map of. Louis J. Moses (in Chapter 1) reminds us of the available theoretical models of mindreading and measures them against each other and against new data. Increasingly, the data point to multiple processes, multiple developmental paths, and a diversity in processes and functions. Bertram F. Malle (Chapter 2) introduces the notion of a *manifold* to characterize the phenomenon of mindreading as a complex array of related but distinct elements. He shows how this idea of a manifold helps resolve three puzzles that any broad theory of mindreading has to solve: how inferences of mind relate to interpretations of behavior, how mental states can be described in language, and how mental state inferences can exist both as automatic, unconscious processes and as deliberate, conscious acts. Mark H. Davis (Chapter 3) presents a component analysis of mindreading by laying out what is currently known about the antecedents, attributes, and consequences of this phenomenon. With this approach, he underscores several questions and problems that have not received adequate attention in the research literature, and this chapter thus helps with the general task of drawing the contours and boundaries of our map. Finally, Daniel D. Hutto (Chapter 4) poses the question of whether the basis of mindreading is grounded in a compact developing framework of concepts (about intention, desire, and belief, for example) or if instead it may initially be nonconceptual. He thus presents a challenge to all major theoretical positions and demands a careful accounting of how the simpler, earlier-developing, and possibly more affective components of mindreading relate to the complex, later-developing, and possibly more deliberate components of mindreading.

One of the unquestioned requirements of mindreading is the perception and analysis of behavior, addressed in Part II. Although people don't have direct access to others' minds, they are typically presented with others' behaviors, which are often reliably correlated with a particular mental state or content. Diego Fernandez-Duque and Jodie A. Baird (Chapter 5) show that such simple tools as following eye gaze are critical to success in mindreading tasks but may be part of domain-general cognitive systems that are useful for many tasks, including those that don't involve mindreading. Susan R. Fussell and her collaborators (Chapter 6) document how these simple tools are used in a social context to take on a larger challenge: regulating collaborative action. Coordination with others requires an appreciation of what information is visually accessible to each party and how it can be used to advance a joint task. However, social interactions often require more than anticipating or guiding someone else's next action. People infer others' goals and plans and evaluate their behaviors. Glenn D. Reeder and David Trafimow (Chapter 7) show the essential function of simple but powerful motive inferences from behavior and how, on that basis, more abstract personality impressions are formed. Stephen J. Read and Lynn C. Miller (Chapter 8) introduce a cognitive model of how these inferences of goals and traits might be integrated in the same system. Such model building that integrates multiple layers of cognitive processing is still rare, but as we learn more about which components are involved in the task of understanding other minds and how they relate to each other, it will be possible to build comprehensive models of social cognition across multiple levels of analysis.

Part III addresses another basic source of information besides a target person's visible behavior: the perceiver's own knowledge, preferences, and experiences. Jean Decety (Chapter 9) suggests that, for some tasks, there is evidence of a processing area in the brain that jointly handles information about self and information about the other and that facilitates some inferences about other minds by using one's own mind as the default. Daniel R. Ames (Chapter 10) examines the process of projecting self-information onto others in the person perception context. He compares projection with the use of general knowledge structures (in this case, stereotypes) and identifies some of the circumstances in which one or the other tool is used. Josef Perner and Anton Kühberger (Chapter 11) go beyond the idea of self as "source of information" to self as an actual simulative process that can deliver the answers to certain question. But, like Ames, these authors view simulation as a useful tool for some, but not all, questions. As simple and appealing as simulation strategies may be, the chapter by Radu J. Bogdan (Chapter 12) analyzes what it actually takes to make inferences of others' mental states on the basis of ascriptions to the self. In doing so, he raises the provocative thesis that inferences about others that are based on the self may actually develop later and be more sophisticated than ones that are not based on the self.

A similar question of developmental sequence arises when considering the relation between mindreading and language, a topic addressed in Part IV of the book. Janet Wilde Astington and Eva Filippova (Chapter 13) make a strong case for a tight interplay between the two capacities: the infant's sensitivity to minds provides a base for language to develop, while participation in conversation and budding syntactic and semantic abilities facilitate mentalistic interpretations of human behavior. The developmental interplay of these two sets of capacities in early childhood sets the stage for interesting questions about their relationship in adulthood. No other social activity requires more negotiation with other minds than conversation. As daunting as these dealings may be, Marjorie Barker and T. Givón (Chapter 14) argue that people's skill at monitoring and adjusting their language in response to their conversation partner may be largely unconscious or, if conscious, then perhaps quickly forgotten. However, it is perhaps precisely because of the unobtrusive nature of these adjustments and the relative ease with which they are made that people may be mindlessly lulled into assuming that their conversation partners share their perceptions and understanding. These assumptions lie at the heart of Michael F. Schober's chapter (Chapter 15), in which he discusses how such assumptions made about ambiguous concepts may have serious consequences not only for interpersonal misunderstandings but also more broadly for society, including misdirection of public policy. James S. Uleman's chapter (Chapter 16) addresses squarely the theme of ambiguity in the meaning of mental concepts. Uleman argues that the very words we use to ascribe mental states and traits to other people are ambiguous and vary greatly across contexts and functions. As Malle discussed in the first section of the book, it is still an open question whether this flexibility is a fundamental limitation of language or perhaps even a necessary ingredient to bridge the diversity of minds.

Flexibility and ease aside, there are contexts in which the limitations, not the marvels, of human mindreading command our attention. And we don't have to stray very far to find examples of these limitations, argue Dale J. Barr and Boaz Keysar in the first chapter of Part V of the book (Chapter 17). As we have seen in the earlier chapter by Schober, normal humans show a tendency to project their assumptions onto others without necessarily considering the *otherness* of that mind. Though these egocentric errors may rarely derail or disrupt everyday interactions or extend to the higher-stakes situations Schober discusses, Barr and Keysar show how they can easily be identified in simple conversations. Barr and Keysar suggest that the egocentric perspective forms a sort of anchor that must be adjusted when considering others' perspectives. This idea is further developed by Leaf Van Boven and George Loewenstein (Chapter 18), who discuss how current bodily states such as hunger and thirst, not just cognitive perspectives, serve as inferential anchors. This tendency to anchor on current visceral states may impede not only people's ability to take the perspective of others but also their ability to consider how they themselves might feel at a different time when their current visceral state has passed. Sara D. Hodges (Chapter 19) illustrates how limitations in mindreading may at times be missed because they are not fully put to the test in social interactions. Assumptions of mutual understanding may sometimes be enough to generate perceptions of sufficient mutual understanding. Whereas Hodges illustrates a case in which inaccurate mindreading may go undetected, William Ickes, Jeffry

A. Simpson, and Minda Oriña (Chapter 20) identify situations in which accuracy may be actively avoided and in which pursuing accuracy could be detrimental. None of these limitations is of the same magnitude, of course, as the ones that Robyn Langdon (Chapter 21) discusses in autistic and schizophrenic individuals. Langdon emphasizes that clinical cases associated with mindreading deficits do not all look alike, and what is missing does not appear to be one single module, system, or capacity. The differential deficits we see in autistic and schizophrenic patients point once more to the fact that mindreading isn't just one thing, that it consists of many tools and elements, and that damage to these potentially independent elements results in quite distinct patterns of deficit.

As excited as we are about a broad approach to the study of other minds and an integration of multiple points of view, this volume is not intended to provide an encyclopedic or exhaustive treatment of the topic of other minds. Instead, we hope that this particular set of insights, with their novel juxtaposition of perspectives, will provide healthy cross-fertilization and creative inspiration to the future study of mindreading. Whereas seemingly intractable problems are things that most people would rather avoid, we anticipate that both ordinary people and scientists will continue to embrace the challenges posed by the problem of other minds.

REFERENCES

Byrne, R. W., & Whiten, A. (1986). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford, UK: Oxford University Press.

Carruthers, P., & Smith, P. K. (Eds.). (1996). *Theories of theories of mind*. New York: Cambridge University Press.

Davies, M., & Stone, T. (Eds.). (1995). *Mental simulation: Evaluations and applications*. Cambridge, MA: Blackwell.

Frith, C. D., & Wolpert, D. M. (Eds.). (2004). *The neuroscience of social interaction: Decoding, imitating, and influencing the actions of others*. Oxford, UK: Oxford University Press.

Frye, D., & Moore, C. (Eds.). (1991). *Children's theories of mind: Mental states and social understanding*. Hillsdale, NJ: Erlbaum.

Malle, B. F., Moses, L. J., & Baldwin, D. A. (Eds.). (2001). *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: MIT Press.

Meltzoff, A. N., & Prinz, W. (Eds.). (2002). *The imitative mind: Development, evolution, and brain bases*. Cambridge, UK: Cambridge University Press.

Zelazo, P. D., Astington, J. W., & Olson, D. R. (Eds.). (1999). *Developing theories of intention: Social understanding and self-control*. Mahwah, NJ: Erlbaum.

This page intentionally left blank

PART I

Questions about the Phenomenon

This page intentionally left blank

1

Executive Functioning and Children's Theories of Mind

LOUIS J. MOSES

The problem of whether we can justifiably be said to know the minds of others is a classic philosophical issue dating back at least to Descartes (1641/1986). In a nutshell, how can we know the unobservable mental states of others when all we have to go on is their external behavior? That is a question for epistemologists. A related but nonetheless conceptually distinct problem concerns how we come to *think* we know other minds. That is, irrespective of whether we are justified in claiming to know other minds, we confidently act much of the time as if we do have such knowledge. How then do we come to make mental state attributions? That is a question for social and cognitive psychologists.

Two readings of this latter question need to be immediately distinguished. One concerns how we make mental state attributions in specific instances. That is, what are the online cognitive processes that underlie particular acts of mental state reasoning? A second reading concerns how we acquire knowledge (or at least beliefs) about mental states in the first place. How do we formulate such abstract concepts as belief, intention, emotion, and the like? This ontogenetic question is the central focus of this essay, although issues of online reasoning will also be addressed.

Several proposals are currently on the table concerning the acquisition of mental state concepts. The first is the so-called theory-theory

(Gopnik & Wellman, 1994), according to which children construct such concepts on the basis of what they observe in the world around them (including whatever they experience of their own inner mental life). The "theories of mind" (ToM) at which children arrive might be largely socially constructed, in which case widespread cultural variation in these theories might be present. Alternatively, they might be heavily constrained by the nature of our cognitive systems and/or by the affordances of the environment, in which case little cultural variation would be expected, at least in core theoretical constructs.

A second approach (simulation theory—see Goldman, 2001) downplays this emphasis on theoretical construction, arguing instead that children have access to their own mental states and then through analogical reasoning and simulation are able to "read off" the mental states of others. According to a third proposal (modularity theory—see Baron-Cohen, 1995; Leslie, 1994), specific neural systems are dedicated to processing social information relevant to mental states. When these systems or modules mature (and when ancillary processing skills are mastered), children are able to make appropriate mental state attributions. A fourth proposal is that the development of mental state concepts is tied to the development of linguistic capacity (deVilliers & deVilliers, 1999). Specifically, children are not able to master a concept such as belief until they have acquired the syntactic frames in which belief utterances are embedded.

In what follows I will lay out the case for a fifth proposal that either competes with or complements these other views. The proposal centers on the role that advances in executive functioning (EF) might play in advances in mental state understanding. Executive functioning is a broad construct encompassing skills and processes such as inhibitory control, planning, set shifting, error detection and correction, and working memory (Welsh, Pennington, & Groisser, 1991; Zelazo, Carter, Reznick, & Frye, 1997). What these abilities have in common is that, in one way or another, they are all implicated in the monitoring and control of action. The possibility that EF might contribute to the development of ToM has received increasing attention in the past decade (Frye, Zelazo, & Palfai, 1995; Hughes, 2002; Moses & Carlson, 2004; Perner & Lang, 1999; Russell, 1996; Schneider, Schumann-Hengsteler, & Sodian, 2005).

IS THERE A RELATION BETWEEN EXECUTIVE FUNCTIONING AND THEORY OF MIND?

There are a number of reasons to suspect that EF and ToM might in some way be linked. The preschool years represent a developmental "sweet spot" with respect to mental state understanding (see Wellman, 2002, for a recent review). At the beginning of this period (2–3 years of age) children appreciate little about mental representations. For example, they generally fail to recognize that beliefs can be false, that appearances may differ from reality, or that different perspectives can be taken on the same scene or event. By the end of that period (5–6 years of age), however, they have an understanding of these matters that in important ways resembles that of adults. Yet, just as these dramatic changes are occurring, so too are there marked advances in children's executive skills (Diamond, 2002; Luria, 1973; Posner & Rothbart, 2000; Zelazo et al., 2003). Throughout this period children become increasingly proficient at pursuing goals in the face of irrelevant distractions, at thinking flexibly about the world around them, and at keeping their behavior and emotions appropriately in check.

Of course, this shared developmental timetable could be entirely coincidental. However, two other sources of evidence suggest that it might not be. First, individuals with autism are known to have profound deficits in ToM (Baron-Cohen, Tager-Flusberg, & Cohen, 2000). It turns out, however, that these individuals also have severe executive deficits (Russell, 1997), suggesting that these two abilities might be connected in some more substantive way. Second, brain imaging studies indicate that adjacent and/or overlapping neural circuits are active when people carry out either EF tasks or ToM tasks (Kain & Perner, 2005). These additional links between EF and ToM are admittedly circumstantial. Nonetheless, they provide a rationale for a direct examination of whether the two constructs are fundamentally related.

In the first major investigation of this possibility, Frye and colleagues (1995) examined whether individual differences in EF and ToM were related. Executive ability was measured using a card sorting task (the Dimensional Change Card Sort) in which children were first required to sort cards by one dimension (e.g., color) and then to sort the very same cards by a conflicting dimension (e.g., shape). ToM was assessed by various measures including false belief and appearance–reality tasks. The false belief tasks included both Wimmer and Perner's (1983) classic location change task and an unexpected contents task (Gopnik & Astington, 1988). In the location change task a protagonist acquires a false belief about the location of an object by virtue of being absent while the object is moved. Children are then questioned about the protagonist's belief (or about where he or she will look for the object). In the unexpected contents task children are presented with a familiar container that turns out to hold something quite different from its usual contents. Children are asked what they initially thought was in the container and/or what a naive individual would think was in it. In the appearance–reality task (Flavell, Flavell, & Green, 1983) children are presented with items that look like one thing but are actually something quite different (e.g., a sponge that looks like a rock). After discovering what the object really is, they are questioned both about what it really is and what it looks like. In several studies using these and similar measures Frye and colleagues found moderately strong correlations between EF and ToM. These findings have since been replicated in numerous studies using a variety of different measures (see Schneider et al., 2005, for recent reviews).

Of course, these data are correlational, and so the observed relations may have been generated spuriously by some other developmental variable: many attributes are related across developmental periods without sharing any deeper connection (e.g., the correlation between height and vocabulary size). Given this lurking danger, it is heartening to discover that when developmental factors are controlled, as well as other variables that are known to relate to either EF or ToM, the relation between the two domains persists. For example, in an extensive study addressing this issue, Carlson and Moses (2001) gave 107 preschoolers a battery of 10 executive measures and another battery of 8 ToM tasks. The two batteries were highly correlated ($r = .66$) and remained so when the following factors were held constant: age, verbal ability, gender, symbolic ability, family size, motor sequencing, and mental state control tasks designed to be structurally similar to ToM tasks but not to require reasoning about mental states (e.g., tasks in which children were questioned about the prior and current location of an object rather than a protagonist's beliefs about the whereabouts of the object). Other studies have found similar results, using quite different measures of verbal ability and/or general cognitive ability (e.g., Carlson, Moses, & Breton, 2002; Davis & Pratt, 1996; Keenan, 1998).

In short, the relation between EF and ToM is not only reliable but also robust. While some irrelevant extraneous variable might yet be found to be at the heart of this relation, many of the obvious candidates have now been examined and ruled out. Hence, any theory hoping to explain the development of children's ToM will need to grapple with this relation. In what follows I evaluate three alternative explanations for why these relations might be found (see Perner and Lang, 1999, for an earlier discussion of these and other possibilities). To foreshadow, I will conclude that the evidence best supports the possibility that EF is necessary for the very emergence of abstract mental state concepts such as belief.

PROPOSAL 1: THEORY OF MIND AFFECTS THE EMERGENCE OF EXECUTIVE SKILLS

The findings discussed so far are all correlational, and so the causal direction between EF and ToM might conceivably run either way. Both possibilities have been argued in the literature. Perner, for example, has argued that the causal power resides largely in ToM (Perner & Lang, 1999; Perner, Lang, & Kloo, 2002). How could ToM influence EF? Perner suggests that metarepresentational capacities at the heart of ToM are also critical for executive skills, such as inhibitory control. Specifically, advances in both ToM and EF require an understanding of the causal power of mental states. In the case of false belief, for example, children need to understand that incorrect information will lead the protagonist to look in the wrong location. In the case of inhibition, children must recognize that an habitual action schema would lead to goal failure, and so must be suppressed in favor of a novel action. Hence, Perner argues, successful executive inhibition, like an appreciation of false belief, depends on metarepresentation: for inhibition, representation of an inappropriate action schema and its relation to action; for false belief, representation of a mental state and its relation to action.

Perner's proposal is a cogent one. Nonetheless, several kinds of empirical data speak against it. For one, longitudinal findings are more consistent with an EF to ToM causal direction. For example, Hughes (1998) found that EF at 3 years of age was a stronger predictor of ToM at age 4 than was ToM at 3 a predictor of EF at 4. In addition, in a younger sample of children, Carlson, Mandell, and Williams (2004) found a similar pattern. EF at 2 years of age significantly predicted ToM at 3 years over and above age, sex, verbal ability, maternal education, and ToM at the earlier age. In contrast, ToM at 2 years of age did not predict later EF over and above controls. Finally, in a microgenetic study of children's false belief performance and inhibitory skills, Flynn, O'Malley, and Wood (2004) found that mastery of inhibitory control developmentally preceded successful false belief performance. Hence, these findings suggest that early EF may be critical for ToM, while the reverse is less apparent.

In addition, training study data more clearly support an EF to ToM account. If the causal direction is from EF to ToM, then training EF skills should generate enhanced ToM performance. Conversely, if the causal direction is from ToM to EF, then ToM training should generate better EF performance. In a training study of this kind (Kloo & Perner, 2003), both hypotheses received some support. EF training led to better EF performance and generalized to enhanced ToM performance. In turn, ToM training led to enhanced EF performance. The latter finding is difficult to interpret, however, because ToM training failed to enhance ToM performance, leaving open the question of what was actually trained. Clearly, we need further studies of this kind to clarify the issues.

On balance, then, while there is admittedly some evidence suggesting a ToM to EF causal direction, the longitudinal and training data more strongly support an EF to ToM causal account. That said, Perner and Lang (1999) point out an interpretive problem with studies of this kind. If, as they argue, EF tasks are actually ToM tasks (because they require metarepresentation), then the fact that performance on EF tasks is a better predictor of performance on ToM tasks over time rather than the other way around is inconclusive. It may just be that what are nominally called EF tasks are, in fact, not only disguised ToM tasks but also more valid measures of ToM than are standard ToM tasks. Similarly, with respect to training studies one could argue that both EF and ToM training are in effect training in metarepresentation, and, hence, data from such studies cannot conclusively resolve the issue of causal direction.

That said, factor analytic evidence to some extent speaks against the idea that EF and ToM tasks are all just measures of metarepresentation. In an analysis of the data reported in Carlson and Moses (2001) a separate ToM factor emerged in addition to the EF factors (Carlson, 1997). If both the EF and ToM tasks were simply measures of ToM (or metarepresentation), one would not expect to see these separate dimensions emerging. Rather, Perner's hypothesis would perhaps predict a single "ToM" factor on which the EF measures would load as highly, or more highly, than the ToM measures.

An additional difficulty for the ToM to EF account lies in the fact that not all inhibitory tasks correlate strongly with ToM. For example, Carlson and Moses (2001) found that two EF dimensions emerged in a factor analysis. They labeled these dimensions conflict and delay. An example of a conflict task is the Bear/Dragon task (Kochanska, Murray, Jacques, Koenig, & Vandegeest, 1996) in which children must respond to the commands of a Bear but not to those of a Dragon. Children often err by following the Dragon's commands as well as those of the Bear. On tasks such as these, children are faced with two conflicting response options and are required to inhibit the prepotent option (following all commands). An example of a delay task is the Gift Delay task (Kochanska et al., 1996) in which children are asked to turn away while the experimenter (noisily) wraps a gift for them. Many children have great difficulty waiting for the requisite period and eventually take a subtle (or not so subtle) peek at the gift. Waiting tasks such as these merely require children to inhibit responding for a certain time. Although both conflict and delay tasks are difficult for preschoolers, only conflict tasks consistently correlate with ToM. In contrast, delay tasks correlate with ToM weakly or not at all (Carlson & Moses, 2001; Carlson et al., 2002; Carlson, Moses, & Claxton, 2004; Hala, Hug, & Henderson, 2003). Yet, on Perner's account, both conflict and delay tasks require metarepresentation and so should correlate about equally with ToM.

A related problem stems from the finding that executive tasks that do relate strongly to ToM require not only inhibitory skill but also working memory. Tasks that make working memory demands but not inhibitory demands, or that make inhibitory demands but not working memory demands, tend not to relate to ToM (Carlson et al., 2002; Hala et al., 2003). However, while Perner's account explains how ToM might be necessary for the development of inhibitory skill, it is not at all clear how advances in ToM could generate working memory advances.

Finally, cross-cultural data are problematic for the ToM to EF proposal. We recently collected data on EF and ToM from a large sample of children in Beijing. Chinese children were of theoretical interest because they live in a culture that places a premium on self-control. We thus expected them to perform better than age-matched North American children on executive tasks. What would then be of interest was whether they would show a similar advantage with respect to ToM. The findings were that the Chinese preschoolers indeed showed an EF advantage but that no such advantage was observed for ToM (Sabbagh, Xu, Carlson, Moses, & Lee, in press). These data are inconsistent with the ToM to EF account. On that account, advances in EF should not be possible in the absence of corresponding advances in ToM. Yet, relative to their North American counterparts, the Chinese children showed advanced EF without an apparent advance in ToM.

PROPOSAL 2: EXECUTIVE FUNCTIONING AFFECTS THE EXPRESSION OF PREEXISTING THEORY-OF-MIND CAPACITIES

Suppose then that the causal direction runs the other way: from EF to ToM. How might that be explained? One possibility is that good executive skills are simply necessary for successful performance on ToM tasks. Indeed a cursory examination of commonly used ToM tasks shows that executive demands are clearly present in those tasks. For instance, in the unexpected location false belief task described earlier, children must hold in mind two representations of the situation—their own presumably salient knowledge of where the desired object is, as well as the protagonist's now outdated representation of the object's location. In addition, children must suppress any inclination to respond in terms of the former and instead reason in terms of the latter. The false belief task thus taxes children's working memory and inhibitory control, two prominent aspects of EF. Similarly, in appearance–reality tasks children must hold in mind both the apparent and real nature of an object, but suppress the latter when questioned about the object's appearance. Hence, ToM tasks such as these require at least two things of children: (1) the relevant ToM concepts and (2) sufficient executive capacity to express these concepts. It is conceivable that children might have the concepts in hand but nonetheless fail ToM tasks because their executive skills are only weakly developed.

The obvious way to test this hypothesis is to manipulate the executive demands of ToM tasks. If children's difficulty is merely one of ToM expression, then such manipulations should systematically affect their task performance. If, on the other hand, their difficulty is primarily conceptual, these manipulations should have little effect. The empirical evidence on this matter is mixed. It is true that *increasing* the executive demands of the false belief task does lead to deterioration in performance in children who would normally succeed (Leslie & Polizzi, 1998). But that is not especially surprising—one could imagine constructing a false belief task that so taxed working memory and inhibition skills that even adults would fail it. The more critical issue is whether *reducing* executive demands uncovers latent conceptual ability in children. In this regard certain kinds of manipulations do indeed generate enhanced performance on false belief tasks. In a meta-analysis of 178 false belief studies, Wellman, Cross, and Watson (2001) uncovered four such manipulations. Children's performance improved if (1) the motive for changing the location of the desired object was deceptive, (2) they were actively involved in changing the object's location, (3) the object was not present at the time the false belief question was asked, and (4) the protagonist's mental state was highlighted by stating or picturing it. All of these manipulations may be thought of as executive because they either downplay the salience of the true state of affairs (3) or increase the salience of the belief (1, 2, 4). Similarly, in our own work (Carlson, Moses, & Hix, 1998) children were more likely to deceive if allowed to use a novel method of deception (e.g., pointing an arrow to "inform" an opponent of a desirable object's location) as opposed to a method that is habitually used to veridically reference the world (pointing with the finger). Successfully deploying the latter method may require considerably greater inhibitory control than that required for making use of a novel response.

Findings of this nature do provide support for the executive expression hypothesis. However, although these manipulations substantially enhance the performance of older preschoolers, they do not move children younger than age 3½ above chance levels. Hence, for younger preschoolers there is little evidence that reducing executive demands uncovers latent conceptual ability (although see Moses, 2001, and Scholl & Leslie, 2001, for some caveats). To the extent then that EF–ToM relations are driven by advances in children's ability to express their ToM, they would not appear to seriously challenge the conceptual deficit explanation of theory-of-mind changes in the preschool period. Younger preschoolers may lack the critical ToM concepts, and their deficits in this regard could well be independent of their executive skills.

PROPOSAL 3: EXECUTIVE FUNCTIONING AFFECTS THE EMERGENCE OF THEORY-OF-MIND CONCEPTS

In addition to influencing the expression of ToM, however, advances in EF might also be critical for the acquisition of ToM concepts themselves (Moses, 2001; Russell, 1996). If so, that would also be a potential explanation for why EF–ToM relations are observed. How could EF play a role in concept acquisition? Imagine a creature entirely devoid of executive skills. Such a creature's behavior would be driven by innate tendencies and learned habits, and would largely be at the mercy of external stimuli. A creature of this kind would be captured by whatever was most salient in the current context and would never be able to distance itself from that context in order to reflect on the possibility of alternative perspectives. Intangible, unobservable entities such as mental states would be inaccessible to it. Some degree of inhibitory capacity would thus seem to be necessary to form mental state concepts (and any other abstract concepts for that matter). Working memory would also seem to be essential. If a creature was never able to hold more than one perspective in mind, it is not clear that one would credit the creature with a notion of perspective at all. In short, some level of executive skill, some reflective capacity, is crucial for forming mental state concepts.

But this proposal seems indisputable: a creature with *no* executive capacity surely can't form abstract concepts. One scarcely needs to collect the data! What does demand an empirical answer, however, is the question of *how much* executive skill is crucial for forming such concepts. After all, we know that well before they begin to succeed on false belief tasks, children have already developed *some* executive capacity, as evidenced, for example, by 1-year-olds' performance on object search tasks (Diamond, 2002). Perhaps this level of executive skill is all that is needed for the emergence of ToM concepts. Whatever EF advances occur in the preschool period could well be irrelevant to concept acquisition.

One way to test this executive emergence hypothesis is to examine children's performance on ToM tasks that require conceptual understanding but that do not impose an undue burden on the executive system. If the executive *expression* hypothesis is correct, then no relation should be found between performance on such tasks and performance on executive tasks—if the tasks are largely free of executive demands, then performance on them should not be associated with children's executive skill. If, on the other hand, the executive *emergence* hypothesis is correct, then a relation should in fact be present. If executive skills are critical for concept formation, then performance on executive measures should predict performance on any task requiring understanding of those concepts, irrespective of whether the task itself makes online executive demands.

Perner and colleagues (2002) carried out a study relevant to this issue. They gave children standard false belief prediction tasks and false belief explanation tasks, and examined how these task types related to performance on executive tasks. As noted earlier, the prediction task carries with it clear executive demands—children must set aside their knowledge of where the desired object is located in order to recognize the protagonist's false belief or to predict where he or she will look for the object. In contrast, in the explanation task there is no prepotent response option needing to be suppressed, and hence the executive demands are minimized. In the explanation task children are simply asked why the protagonist looked in the incorrect location (the correct answer being that he or she held a false belief). Consistent with the executive emergence hypothesis, Perner et al. found that performance on the executive tasks was just as highly correlated with performance on the explanation task as with performance on the prediction task.

In our own work the same pattern was found using two quite different tasks that, like the explanation task, do not include prepotent response options (Moses, Carlson, Stieglitz, & Claxton, 2005). In a mental state certainty task (Moore, Pure, & Furrow, 1990), an object was hidden in one of two locations, and children then listened to the advice of two puppets. One puppet stated that it knew the object was in a particular location, while the other stated that it merely thought the object was in the other location. The children's job was to find the object. The task thus tests children's knowledge of the distinction between "know" and "think." This is a difficult task for preschoolers, and performance on it is correlated with false belief performance (Moore et al., 1990). However, although children frequently err, they do not do so systematically—across trials, performance is more or less random. In other words, unlike the actual location of the desired object in a false belief task, the incorrect location in the mental state certainty task is not prepotent. Hence, the task would seem to impose few executive demands. The second task of this nature was the sources-of-knowledge task (O'Neill & Gopnik, 1991). In this task children learn the identity of an object either by touching it, seeing it, or hearing about it. They are then asked how they knew the identity. This task is difficult for preschoolers, but, again, their errors are unsystematic: they do not, for example, show a clear preference across trials for claiming to have attained knowledge through seeing.

We assessed children on these tasks as well as executive tasks and standard ToM tasks, like the false belief task. The findings were again inconsistent with the executive expression hypothesis—the executive tasks correlated just as highly with these nonprepotent ToM tasks as with the standard prepotent tasks. These results support an executive emergence account. Even when there are no obvious executive demands, EF–ToM relations persist as would be expected if EF were implicated in the formation of ToM concepts themselves.

Further evidence consistent with the emergence hypothesis but problematic for the expression hypothesis comes from Carlson and colleagues' (2004) longitudinal data. The authors found that EF at 2 years of age predicted ToM at age 3 over EF at age 3. However, if children's difficulty were one of expressing existing knowledge, then it is difficult to explain how executive skills that have already developed (those measured at age 2) could influence ToM performance over and above those that are in the process of developing (those measured concurrently at age 3). In contrast, if EF affects the acquisition of ToM concepts, then EF skills emerging just prior to the advent of ToM skills (those measured at age 2) might well contribute additional predictive power. And the latter is what Carlson and colleagues' data suggest.

Finally, the cross-cultural data described earlier (Sabbagh et al., in press) are also inconsistent with the ToM expression account. Recall that, relative to North American children, Chinese children were advanced on EF but not on ToM. These children appeared to have an abundance of executive capacity, and yet the presence of this capacity did not lead them to reveal hidden ToM competence. The findings are, however, compatible with the executive emergence account. On this account, executive advances are necessary for the emergence of ToM concepts. They are not, however, sufficient: executive skills may make possible the emergence of ToM concepts, but they cannot create those concepts. Hence, on the emergence account an advance in EF would not necessarily be immediately matched by a corresponding advance in ToM.

In sum, the existing data seriously challenge the expression hypothesis. That hypothesis is unable to explain the full range of correlational findings. Those findings can however be readily accommodated within an executive emergence framework.

CONCLUSION

Reasoning about the minds of others is a complex, multifaceted skill requiring both domain-specific and domain-general abilities. The evidence reviewed here suggests that one domain-general ability—executive functioning—is intimately bound together with children's developing theories of mind. Relations between EF and ToM in the preschool period are both reliable and robust. These relations may be partially explained by the role that executive advances play in the expression of children's ToM and by the role that ToM advances play in the development of executive skill. At this point in time, however, the overwhelming share of the evidence is most consistent with the hypothesis that executive skills play a crucial role in the very emergence of children's theories of mind.

At the outset of this essay I listed a number of potential explanations for the marked changes in ToM that occur during the preschool years. If the executive emergence hypothesis is indeed correct, what are the implications for these other explanations? The hypothesis complements both the "theory-theory" and simulation theory. With respect to the theory-theory, it may well be that changes in children's theories of mind are in some ways akin to scientific theory change. However, executive advances may be critical in making such changes possible. Similarly, executive advances could be crucial both for the growth of reflection and the development of analogical reasoning, upon which simulation theory depends.

In contrast, the emergence hypothesis conflicts with both modularity theory and syntactic theory. According to modularity theory, developments in ToM occur on a maturational timetable, and the proposed ToM modules should be cognitively impenetrable by domain-general processes such as executive functioning. Similarly, the acquisition of syntax is traditionally viewed as a highly domain-specific process, and so a role for domain-general processes such as EF could not be accommodated within the syntactic theory of ToM without sweeping revisions of that theory. In any event, it is clear that all accounts of theory-of-mind development will need to grapple more seriously with the role that domain-general abilities such as executive function undoubtedly play in children's developing appreciation of mental life.

ACKNOWLEDGMENTS

Thanks to Sara Hodges for constructive comments on an earlier draft of this chapter.

REFERENCES

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. J. (Eds.). (2000). *Understanding other minds* (2nd ed.). New York: Oxford University Press.

Carlson, S. M. (1997). *Individual differences in inhibitory control and children's theory of mind*. Unpublished doctoral dissertation, University of Oregon.

Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: Stability and prediction from age 2 to 3 years. *Developmental Psychology, 40*, 1105–1122.

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

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*, *11*, 73–92.

Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Executive function and theory of mind: The role of inhibitory control and planning ability. *Journal of Experimental Child Psychology*, *87*, 299–319.

Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory control in young children's difficulties with deception and false belief. *Child Development, 69*, 672–291.

Davis, H. L., & Pratt, C. (1996). The development of children's theory of mind: The working memory explanation. *Australian Journal of Psychology*, *47*, 25–31.

Descartes, R. (1986). *Meditations on first philosophy.* New York: Cambridge University Press. (Original work published 1641)

DeVilliers, J. G., & deVilliers. P. A. (1999). Linguistic determinism and false belief. In P. Mitchell & K. Riggs (Eds.), *Children's reasoning about the mind*. Hove, UK: Psychology Press.

Diamond, A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D. Stuss & R. Knight (Eds.), *Principles of frontal lobe function* (pp. 466–503). New York: Oxford University Press.

Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance–reality distinction. *Cognitive Psychology, 15*, 95–120.

Flynn, E., O'Malley, C., & Wood, D. (2004). A longitudinal, microgenetic study of the emergence of false belief understanding and inhibition skills. *Developmental Science*, *7*, 103–115.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development, 10*, 483–527.

Goldman, A. I. (2001). Desire, intention, and the simulation theory. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 207–224). Cambridge, MA: MIT Press.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance–reality distinction. *Child Development, 59*, 26–37.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press.

Hala, S., Hug, S., & Henderson, A. (2003). Executive functioning and false belief understanding in preschool children: Two tasks are harder than one. *Journal of Cognition and Development*, *4*, 275–298.

Hughes, C. (1998). Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind? *Developmental Psychology*, *34*, 1326–1339.

Hughes, C. (2002). Executive functions and development: Emerging themes. *Infant and Child Development*, *11*, 201–209.

Kain, W., & Perner, J. (2005). What fMRI can tell us about the ToM–EF connection. In W. Schneider, R. Schumann-Hengsteler, & B. Sodian (Eds.), *Young children's cognitive development: Interrelationships among executive functioning, working memory, verbal ability, and theory of mind* (pp. 189–217). Mahwah, NJ: Erlbaum.

Keenan, T. (1998). Memory span as a predictor of false belief understanding. *New Zealand Journal of Psychology, 27*, 36–43.

Kloo, D., & Perner, J. (2003). Training transfer between card sorting and false belief understanding: Helping children apply conflicting descriptions. *Child Development*, *74*, 1823–1839.

Kochanska, G., Murray, K., Jacques, T. Y., Koenig, A. L., & Vandegeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development, 67*, 490–507.

Leslie, A. M. (1994). ToMM, ToBY, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). New York: Cambridge University Press.

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science, 1*, 247–253.

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology.* New York: Basic Books.

Moore, C., Pure, K., & Furrow, D. (1990). Children's understanding of the modal expression of certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development, 61*, 722–730.

Moses, L. J. (2001). Executive accounts of theory of mind development. *Child Development, 72*, 688–690.

Moses, L. J., & Carlson, S. M. (2004). Self regulation and children's theories of mind. In C. Lightfoot, C. Lalonde, & M. J. Chandler (Eds.), *Changing conceptions of psychological life* (pp. 127–146). Mahwah, NJ: Erlbaum.

Moses, L. J., Carlson, S. M., Stieglitz, S., & Claxton, L. J. (2005). *Executive function, prepotency, and children's theories of mind*. Manuscript in preparation, University of Oregon.

O'Neill, D. K., & Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology, 27*, 390–397.

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Science*s, *3*, 337–344.

Perner, J., Lang, B., & Kloo, D. (2002). Theory of mind and self control: More than a common problem of inhibition. *Child Development*, *73*, 752–767.

Posner, M. I., & Rothbart, M. K. (2000). Developing mechanisms of self regulation. *Development and Psychopathology, 12,* 427–441.

Russell, J. (1996). *Agency: Its role in mental development*. Hove, UK: Erlbaum.

Russell, J. (Ed.). (1997). *Autism as an executive disorder.* New York: Oxford University Press.

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (in press). The development of executive functioning and theory of mind in young preschoolers from Beijing, China: A cross-cultural approach. *Psychological Science*.

Schneider, W., Schumann-Hengsteler, R., & Sodian, B. (Eds.). (2005). *Young children's cognitive development: Interrelationships among executive functioning, working memory, verbal ability, and theory of mind*. Mahwah, NJ: Erlbaum.

Scholl, B. J., & Leslie, A. M. (2001). Minds, modules, and meta-analysis. *Child Development*, *72*, 696–701.

Wellman, H. M. (2002) Understanding the psychological world: Developing a theory of mind. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 167–187). Malden, MA: Blackwell.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development*, 72, 655–684.

Welsh, M. C., Pennington, B. F., & Groisser, D. B. (1991). A normative-developmental study of executive function: A window on prefrontal function in children. *Developmental Neuropsychology*, *7*, 131–149.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128.

Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of General Psychology*, *1*, 1–29.

Zelazo, P. D., Müeller, U., Frye, D., Marcovitch, S., Argitis, G., Boseovski, J., et al. (2003). The development of executive function in early childhood. *Monographs of the Society for Research in Child Development*, *68*, 3(Serial No. 274).

2

Three Puzzles of Mindreading

BERTRAM F. MALLE

Mindreading is the human activity of inferring other people's mental states. Literatures from multiple disciplines have advanced our knowledge on this phenomenon, and we now know quite a bit about the development and functional use of mindreading in the human species (e.g., Astington, 1993; Baron-Cohen, Tager-Flusberg, & Cohen, 2000; Malle, Moses, & Baldwin, 2001; Perner, 1991; Wellman, 1990) and are even beginning to sketch a picture of its evolutionary origins (Baron-Cohen, 1999; Bogdan, 2000; Malle, 2002; Povinelli, 2001; Whiten, 1999). But as we try to integrate this growing knowledge and perhaps work toward a unified theory of mental state inference, many problems and puzzles emerge. In this chapter I focus on three such puzzles. Each of them, I will argue, has a credible solution, but what is perhaps more important is that each of these solutions points to the same general conclusion about mindreading.

THE FIRST PUZZLE: BEHAVIOR AS INPUTS AND OUTPUTS OF MINDREADING

The first puzzle concerns the relationship of behavior to inferences of mental states. On the one hand, most researchers assume (or explicitly theorize) that people typically infer mental states by observing the actor's behavior, making perceptions of behavior an important input to mindreading. Watching facial expressions, head and eye direction, body posture, and movements allows the perceiver to figure out a great deal about what others believe, want, feel, or intend. On the other hand, perceptions of behavior can also be an output of mindreading. Plenty of research shows that people explain behavior to a considerable extent by ascribing mental states to the agent (e.g., Buss, 1978; Heider, 1958; Malle, 1999; McClure, 2002; Read, 1987; Wellman, Hickling, & Schult, 1997). In the case of intentional behavior, it is primarily beliefs and desires that are seen as reasons that make the agent's action intelligible. But even in the case of unintentional behavior, mental states such as emotions, perceptions, and thoughts serve to explain why the person behaved in the observed way. Thus, mental state inferences are inputs to explaining behavior.

But if this is all true, we are caught in circularity. How can the human perceiver use an agent's observed behavior to infer his or her mental states but gather the meaning of that very behavior from inferences of mental states? We know there has got to be a solution to this puzzle; but how do people get it done?

There would seem to be two paths out of this circularity. One path requires at least some behaviors whose meaning can be assessed without reference to mental states. Those behaviors could then be noncircular inputs to certain mental state inferences. The other path requires that at least some mental states can be grasped without behavior observation, and those mental states could then be noncircular inputs to the interpretation of certain behaviors.

Along the first path, I will first discuss two less promising approaches, followed by two more promising ones.

Raw Behaviors

One possibility is that perceivers use "raw" behavior observations as input to mindreading—behaviors that don't require any further interpretation. But raw behavior observations are hard to come by. Purely physical descriptions of an agent's body (without making any reference to meaning) are difficult to produce unless one is an exercise and movement scientist. Neither young children (past the age of 4, when they make a variety of mental state inferences) nor untrained adults have such technical knowledge. Moreover, minute variations in physical behavior would demand distinct physical descriptions. These descriptions would somehow have to be translated into *types* of indicator behavior, because only then would the perceiver be able to narrow down the possible mental states to infer. Without at least a rudimentary meaning analysis, however, such type identification would seem extremely difficult.

In fact, a purely physical analysis of behavior resembles what we see in autistic individuals—who find both social behavior and mental states incomprehensible. Autistic perceivers apparently notice "raw" behaviors but do so without recognizing their meaning. One autistic person reports:

> I know people's faces down to the acne scars on the left corners of their chins and what their eyes do when they speak, and how the hairs of their eyebrows curl, and how their hairlines curve around the tops of their foreheads. . . . The best I can do is start picking up bits of data during my encounter with them because there's not much else I can do. It's pretty tiring, though, and explains something of why social situations are so draining for me. . . . That said, I'm not sure what kind of information about them I'm attempting to process. (Blackburn, Gottschewski, George, & L—, 2000)

Some other information sources have to supplement the raw perception of behavior for those behaviors to appear meaningful. Perhaps perceivers can use context information to make a type identification. Tears flowing down the agent's face during a funeral indicate a different behavior than tears flowing down the agent's face after listening to a joke. That would distinguish crying from laughing, and once these types are distinguished, they could serve as inputs to mental state inferences of sadness and joy, respectively. But what exactly is "context"? It isn't just the physical arrangement of bodies and props; it is the interpretation of those physical elements as *standing for* something, *counting as* a funeral. It would be difficult to classify most contexts without registering what the participants take the situation to be—requiring inferences about their beliefs, assumptions, and interpretations (Givón, 2005), and throwing us back into circularity.

Intentionality Concept

A more promising approach is to consider the concept of intentionality—a conceptual frame within which behavior is interpreted but, at least among infants, without the inference of mental states. Infants learn early to distinguish intentional from unintentional behavior, with estimates ranging from 9 months to 18 months (Carpenter, Akhtar, & Tomasello, 1998; Gergely, Nádasdy, Csibra, & Bíró, 1995; Woodward, 1999). At this tender age, infants must respond to certain cues: the degree of movement "smoothness" as a symptom of the agent's control (compare walking down stairs versus falling down stairs); characteristic accompanying behaviors (e.g., head turning, eye gaze); the connection and manipulation of objects in the world; equifinality (the principle that intentional agents pursue their goals along multiple paths, trying a different path when the first one failed); and characteristic agent responses at the end of the behavior (e.g., "there!" or a happy face with an intentional action; "oops" or an unhappy face with an unintentional action). As far as we know, intentionality is not "mentalized" at this early age; rather, the concept serves as a means of distinguishing observable behaviors into two classes, with room for further differentiation within each class.

With increasing age (probably between 2 and 4 years), children learn to judge intentionality with more sophistication and thus begin to appreciate the involvement of mental states in action—states such as desires, beliefs, and intentions. More specifically, children slowly learn the adult concept of intentionality, which incorporates four mental states. An action is considered intentional only when the agent has desire for an outcome, beliefs about the action that leads to the outcome, an intention to perform the action, and awareness of fulfilling this intention while acting (Malle & Knobe, 1997a). This frame allows the powerful inference that whenever an action is intentional (presumably judged, like the infant does, on the basis of cues other than mental states), there must be the involvement of various mental states, most notably beliefs and desires that provide the reasons for acting (Malle, 1999). Obviously, the conceptual frame does not provide context-specific beliefs and desires—the perceiver knows only that *some* desire and *some* beliefs were involved, not which ones (Malle et al., 2001).

Intentionality thus provides an interpretational frame that even toddlers can acquire before they infer any mental states. Once this frame is in place, expectations for and responses to intentional behavior can be better coordinated, and the significant elements of a social interaction appear in relief. Moreover, with time, the intentionality concept facilitates the inference of mental states in a noncircular way, using the "trick" of postulating certain *kinds* of mental state upon encountering intentional behavior—a postulate that holds without any further analysis of the behavior's specific meaning.

Transparent Behaviors

Another promising approach along the first path is the identification of behaviors that may be sufficiently transparent so that the perceiver who uncovers their meaning gets by without any mental state inferences. The first such class is expressive behaviors, such as screaming in pain, laughing with joy, or growling with anger. Their meaning could initially be purely functional in that they are associated with certain antecedents (e.g., in the case of screaming in pain, sharp or heavy objects intruding on the body) or certain consequences (e.g., in the case of growling with anger, destructive movements). These antecedents and consequences are part of the *physical* context and therefore require no mental state interpretation. The second class of transparent behavior comprises basic human movements such as reaching, grasping, walking, standing up, and lying down. The meaning of such actions can initially be purely functional as well—defined by the role they play in interactions with objects or other beings. For example, reaching and grasping connect the agent with objects and make those objects manipulable and consumable.

In addition to their transparency, expressive behaviors and basic movements have another important feature: They are performed early by human infants themselves. This feature has two associated characteristics. First, transparent behaviors are also the kinds of behaviors that have a high probability of being imitated or becoming "contagious" (see next section), and at least some of them have been linked to brain structures that translate perceptions of another person's behavior into motor programs of performing that same behavior (Decety, Chapter 9, this volume; Jeannerod, 1994; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). Second, behaviors that the infant performs will, from an early age on, be associated with an "inside" perspective—the experience of performing those behaviors in context (Russell, 1996). This experiential perspective may well be one starting point for inferring experiences in the *other* person who is observed performing those behaviors. First a link is created between representations of behaviors (e.g., seeing a growling face; observing a reach-and-grasp movement) with one's own characteristic experiences when performing these behaviors (e.g., feeling angry and agitated; succeeding in pursuit of a goal); later, this associative link between behavior representation and first-person experience may be "transferred" such that observing another's behavior does not just help simulate first-person experiences in the perceiver but triggers the third-person inference that the other has those characteristic experiences as well (Goldman, 2001; Meltzoff & Brooks, 2001).

In sum, expressive behaviors and fundamental movements are embedded functionally in physical contexts, which makes them transparent and thus excellent candidates for being comprehended without any mental state inference. Moreover, transparent behaviors allow for a link between representation of the other and experience of oneself, making them excellent candidates for facilitating mental state inferences.

Effect States

I now turn to the second path of resolving our puzzle—this one requiring mental state inferences that do not rely on behavior input. Such inferences target one of two very different classes of mental states. Let us roughly call them mental states that are often *effects* and mental states that are often *causes* of behavior. Those that are effects include perceptions and emotions. They can be predicted fairly well from the context, general knowledge of the actor, one's own reactions (simulated or shared), and a set of lawlike generalizations about what people want, need, see, think, and feel under a variety of conditions. These mental states are relatively easy to infer because they constitute relatively reliable responses to the world. Now, when such states are associated with expressive behaviors (e.g., eye movements, facial expressions), we return to the first path of solving the puzzle. But these "effect" states may well be inferred even before they are expressed, and even in situations that do not allow for observation of the agent's expressive behavior (e.g., when the perceiver is not physically copresent with the agent).

By contrast, there are mental states that exist primarily as causes of behaviors (or of other mental states), such as desires, beliefs, and intentions. These states are far more difficult to infer, because they are context-specific and can be quite idiosyncratic. Their inference will rely on transparent behavior input (expressive behaviors, basic movements) and additional knowledge about the agent's past actions and general dispositions. For example, when an agent revealed such a mental state in past circumstances (e.g., when explaining his or her action), the ascribed desires and beliefs can be stored and held ready for the next time that certain identical or parallel triggers in the environment are observed to be present. Some form of intentional perspective taking may also be recruited for these purposes ("In his position, what would I want, like, or think?"). So, whereas *effect* states may be inferable without behavior input, *cause* states require behavior input as well as several other sources of information.

We can now summarize the likely ways in which human perceivers solve the inference puzzle that emerges between behavior and mental states. Along one path, they are sensitive to certain "transparent" behaviors that do not require prior inferences of mental states. In addition, they are sensitive to the intentionality of behaviors (judged from cues other than mental states), which in turn guides the selective search for mental states and thus kick-starts the social-cognitive system without circularity. Along a second path, human perceivers are sensitive to the difference between *effect* mental states and *cause* mental states, with the former being to some degree predictable or inferable from physical contexts or transparent behaviors, none of which require a meaning analysis that itself presupposes mental states. The human perceiver is thus not operating circularly but rather by using a parallel strategy: identifying behaviors whose meaning can be understood without prior analysis of mental states and identifying mental states whose existence and meaning can be established without prior meaning analysis of behavior. These kinds of behaviors and these kinds of mental states provide starting material for the social-cognitive system, and once it gets started, it can easily ratchet up its inferences, because starting behaviors can then be used for inferring more difficult mental states, and starting mental states can be used to explain the meaning of more difficult behaviors.

THE SECOND PUZZLE: HOW DO WE SPEAK ABOUT THE MIND?

The second puzzle concerns the ways human perceivers represent and identify mental states in language. It seems clear that humans talk about the mind. In doing so, they are not deterred by the fact that mental states are unobservable; after all, there are plenty of unobservable states and objects that we comfortably discuss (e.g., justice, electrons, the Big Bang). But the exact relationship between mental states and the words that refer to them is quite unclear. On the one hand, mental state terms are notoriously difficult to define, certainly vague, sometimes equivocal, as Uleman (Chapter 16, this volume) argues. Many emotion words, for example, refer to behaviors (a sad look, an angry face) as much as to internal states (I feel sad, she is very angry); it would take us a while to explain what it means to *trust* someone; *believing* can refer to a spiritual, intellectual, perceptual, or emotional internal state; waking up from our dreams, we struggle to find words that describe the feelings, thoughts, and images we just inhabited. One way to summarize these mind–language relations is to propose that there are a limited number of language terms that are used for a large variety of mental states, acts, and experiences, changing from context to context, from agent to agent, thereby blurring the boundaries of meaning.

But others diagnose the exact opposite problem. Recently, Sabini and Silver (2005) argued that the language of mental states is far richer than the mental world they describe, that is, "There are fewer unique mental states than one might have thought" (p. 9). This is a rather counterintuitive position, so let's look closely at some of the evidence that Sabini and Silver provide for their claim.

In a vignette study (Sabini, Garvey, & Hall, 2001), participants imagined being the protagonist in a story in which someone was helping them move from one office to another. In the process, the helper stumbles on a piece of pornography that either (1) truly belongs to the protagonist or (2) actually belongs to a former office occupant—though the helper doesn't know that. In case (1) participants described their emotion as equal parts of shame and embarrassment, whereas in case (2) they described it more as embarrassment than shame. The authors took this to be evidence that "shame and embarrassment are different interpretations of the same raw feeling and these interpretations take into account the different conditions that surround them" (p. 6). But what is this "same raw feeling"? Who makes the judgment of sameness?

The protagonist's psychological world is, in each case, an interwoven complex of realizing, fearing, sensing, intending, and much more; and there is clearly some overlap of these complexes across the two cases. But how do we pick out exactly one "raw feeling"? Sabini and Silver (2005) treat the mental world as segmentable and its segments as uniquely identifiable, but that is a treatment best applied to physical objects and not psychological states. Moreover, the complex psychological state each protagonist experiences goes far beyond what the words *embarrassed* or *ashamed* can capture. The vignette in fact may illustrate the limits of language to reference complex mental states rather than an abundance of language for the "same" mental states.

In another study (Silver & Sabini, 1978), the researchers created a video in which a (male) student tells a (female) friend that he didn't get into medical school. Soon thereafter, a mutual friend enters and excitedly reports that he got into Harvard Medical School. On further questions by the female friend, he also mentions that he received a full scholarship. After the Harvard admittee leaves, the unsuccessful student complains to the female that their friend was bragging. Participants who saw this video are asked to describe the complaining student's emotion, and they overwhelmingly ascribe envy to him. Arguably, however, the emotion that the complaining student himself felt was righteous indignation or anger. From this Sabini and Silver (2005) conclude that "the experiences of an envious person and of a righteously indignant person can be the same" (p. 15), that "righteous indignation and envy . . . are the same experiential state" (p. 15). Thus, Sabini and Silver take this to be another illustrative case of two language labels referring to one and the same emotional state.

But do spectators and the actor himself refer to the *same* emotional states? Again, who would be the judge of this sameness? The actor would certainly deny that he feels envy—and, if we let him be perfectly honest and truthful, he will not *feel* envy (at least in the moment). So how can all spectators agree that his emotion was one of envy, not of anger? Are they right in their third-person perspective, and is the actor wrong in his first-person perspective?

I want to suggest that in both stories (the pornography discovery and the medical school announcement), we are dealing with vaguely bounded *emotion–action complexes*, and different descriptors pick out different aspects of those complexes. Take the medical school story. By stipulation, the actor describes what he is consciously feeling at the moment; but the spectators describe the whole emotion-plus-action complex with which they are presented (reflecting a rather general actor–observer asymmetry in people's attention to experiences vs. actions; Malle & Knobe, 1997b; Malle & Pearce, 2001). No spectator would censure the actor by saying "No, no, you *feel envious*!" At best, they might say "You *are* envious," describing not a conscious experiential state but a combination of antecedents, behaviors, and unconscious emotions. Consider this parallel case: If a spectator said "He is envious" and the actor said (honestly) "I am not envious," we would not conclude that *envious* and *not envious* are two terms for the same state. Nor should we conclude, as Sabini and Silver (2005) suggest, that the terms *envy* and *righteous indignation* refer to the same single state. All that we should conclude in each case is that the parties are talking about two different things.

Likewise in the discovery story: The very fact that participants described the imagined feeling using both terms of shame and embarrassment, but in slightly different blending, suggests that we are dealing with complex and subtly different experiences. There are of course some similarities in the two variants of the story. In each, the protagonist realizes that the helper just found a piece of pornography and that the helper will assume it is the protagonist's. But then the differences begin. The one who actually owns the piece may feel "caught" and as a result be mortified; the nonowner may be surprised, perhaps shocked, fearing that the helper will have a (false) bad impression and quickly searching for ways to change the helper's impression. I have already used more than half a dozen mental state descriptors to describe these complex experiences, and I have not even begun to capture their nuances.

Emotions and experiences normally don't have boundaries that would allow us to reliably identify or count them like pebbles on the beach or words in conversation. Any given emotion (or, more generally, every state of mind) is a complex of combined and recombined mental and physical states, often tied up with intentions and actions. The corresponding language of those states is similarly complex in that it consists of terms that individually have context-sensitive meaning and are flexibly combined and recombined to represent the complexity of the mental world. And that is why, as mentioned earlier, mental language is vague and hard to define. Vagueness is the very feature that allows a limited number of terms to begin to describe a far larger (arguably infinite) number of states.

But how, one might object, could we ever use such a vague language of mind in a consistent, publicly shared way? In particular, how can observers even approximate an actor's complex experience with a handful of vague terms and limited behavioral evidence? The solution lies in a diversity of criteria that allow people—actors, interaction partners, or spectators—to assess the appropriateness of a mental state term used in a particular context. These criteria include (but are probably not limited to) introspection, memory, observation, joint attention, logic, and negotiation. Any given mental state ascription may not draw on evidence from all criteria, but as long as the available evidence reasonably converges, there will be stability in talk about the mind.

Scholars are often nervous about the role of introspection as a criterion of mental state ascriptions. Some (typically dismissed as "Cartesians") see it as the fundamental criterion against which others have to be measured; others see it as an illusion (e.g., Ryle, 1949). Wittgenstein (1953) argued that actors cannot use introspection as the criterion of their mental state talk, because it would mean speaking a "private language," which is no language at all. But there is no denying that actors sometimes do use their conscious experience as the guide to choose words of mind—for example, in response to legitimate questions such as "What are you thinking right now?" or "What are you feeling right now?" Actors just cannot always and solely rely on introspection to describe their psychological states; considerations of what others know, observe, remember from the past, and assumptions of logic and plausibility will often figure prominently as well. Similarly, observers will sometimes rely entirely on the actor's self-report to learn about another's mind—which is precisely why they ask questions such as "What are you thinking right now?" or "What are you feeling right now?" But observers do not always confine themselves to that kind of evidence; they may have reason to doubt the actor's self-report, or they may be interested in a more complex psychological state that is not solely constituted by conscious experience (as in the ascription of *envy* to the unsuccessful medical school applicant).

The meaning of mental state terms is thus not restricted to a kind of "private pointing" to inner states. Once we give up the idea that mental state terms rigidly refer to precisely bounded states, the use of any mental state term becomes a social act, and the appropriateness of this use is subject to the full variety of criteria available. To make a justifiable mental state ascription observers need to take into account what they see and know about the agent; what they generally know about people and the type of context the agent is in; what the agent reports, and how much he or she can be trusted; what others say or would say; and what the goals and stakes are of everyone involved. Precisely because observers make use of all this evidence, actors, too, must take it into account if they want to be credible to observers or converge with their judgments.

There will be cases of disagreement between actors and observers (and the medical school discovery story illustrates one of them). Facing such disagreement, people will clarify and compare their evidence, negotiate the relevance of this evidence for the claim at issue, and occasionally conclude that they were talking about different things. Once more, the vagueness of mental state terms is not a nuisance here but a blessing—it is the feature that helps bridge the actor–observer gap. Only because these terms are vague and often refer to an unspecified mix of internal states, behavioral indicators, and contextual constraints can actors and observers find sufficient agreement to render the language of mind intersubjective and meaningful. It may, in fact, be a necessary requirement of a language of mind to have unsharp boundaries and meanings that vary with context, neighboring terms, and the interlocutors' goals.

THE THIRD PUZZLE: IS MINDREADING A HIGH-LEVEL OR LOW-LEVEL ACTIVITY?

The final puzzle is that, even though mindreading appears to be a sophisticated and challenging activity of higher cognition, much of the mindreading that goes on in everyday interaction is not conscious, and some may not even be cognitive. I consider two cases here, one involving physiology in interaction, the other involving tacit inferences in real-time conversation.

When two people interact, their bodies often begin to synchronize in a number of ways: in posture, gesture, facial expression, timing and structure of speech, heart rate, and more (for reviews, see Chartrand, Maddux, & Lakin, 2005; Levenson & Ruef, 1997). Such synchronization will be imperfect and can at times be entirely absent; but often it is remarkable and has led to compelling demonstrations of behavioral mimicry (e.g., Bavelas, Black, Lemery, & Mullett, 1986; Bernieri, 1988; Chartrand & Bargh, 1999) and physiological linkage (Levenson & Ruef, 1992). Similar mechanisms are also responsible for the phenomenon of emotional contagion (Hatfield, Cacioppo, & Rapson, 1994), detectable in newborns who begin to cry when they hear other babies cry (Simner, 1971), in adults who quietly sit together and adopt one another's moods (Friedman & Riggio, 1981), and in crowds that may break out in violence once a few individuals model violence (Patten, 1999).

The best explanation for synchronization in interaction is that one person's emotional or physiological state is expressed in his or her behavior, this behavior is automatically imitated by the other person, in whom (on the basis of well-practiced associations and perhaps a common neural coding system, Jeannerod & Frak, 1999; Meltzoff, 2002) similar internal states are generated. If needed, the second person could then correctly represent the first person's emotional or bodily state merely by relying on the default assumption that, all else equal, others will be in a similar state as he himself or she herself is. In this case at least, the assumption leads to accurate judgments because the perceiver's "evidence" for the other's mental state was caused by that very mental state (mediated by expressing and imitating bodies). Such active representation, however, will often be superfluous; the very synchronization may suffice to guide the interaction, leading to rapport and cooperation (Barsade, 2002; Bernieri, 1988). In a sense, mindreading has occurred without anybody really trying to read the other's mind.

The second case comes from conversation. Having a successful conversation requires a person to continuously track the other person's beliefs, goals, intentions, and emotional reactions. Some of this tracking consists of explicit perspective taking (conscious and deliberate reasoning about the other mind), such as when the pining teenager sits next to the class beauty and wonders whether she likes him. Some of the tracking occurs unconsciously but results in the conscious ascription of a specific mental state, as when it dawns on you that a friendly stranger on the street actually wants to sell you something; or that your politely nodding colleague doesn't actually know what the acronym HLM stands for. But, as Barker and Givón (Chapter 14, this volume) argue, much of this tracking in ordinary conversation occurs unconsciously. There are uncountable examples of speakers adjusting word choice and grammatical forms depending on what they think the audience understands (Fussell & Krauss, 1992; Givón, 2005). Here is one: "When you are ready to leave, just knock on *the door*," the homeowner says, disappearing from the repair person's view behind a sliding door. The homeowner can say "the door" because she can safely assume that the repair person knows which door to knock on—the one behind which she disappears while the repair person is watching. If she had called out that utterance from the other side of the house, she would have had to specify something like "the sliding door at the end of the hallway."

In general, acts of reference (to objects, actions, locations, times, etc.) are subtle exercises in perspective taking, as the speaker must consider what the other knows. Now, we shouldn't expect that speakers do some sort of calculations in every case. In fact, there are at least three sources of information that might preempt actual mental state inferences.

First, one's own perspective can serve as the standard. For example, if I recall in my conversation that I previously mentioned that my wife's name is Lara, I will subsequently refer to her simply by her name, Lara, implicitly assuming that my conversation partner also remembers that I said who “Lara” is. These kinds of situations are of course open to error—because it is so natural for me that the name *Lara* refers to my wife, it may not even occur to me that my current conversation partner does not know that. (For further examples and evidence of such errors, see Barr & Keysar, Chapter 17, this volume.)

A second preemptive source of information is generic knowledge—knowledge that normally all members of a given community share (e.g., “*The* sun is coming out” vs. “*A* storm is rolling in). This will predictably get you into trouble when you misjudge your conversation partners’ community membership (Fussell & Krauss, 1992).

Third, well-practiced scripts can support an action that seems to imply a certain mental state inference but is not actually made, at least not consciously. For example, after taking the order, the waiter reaches his hand out toward the guest, “assuming” that the guest knows what the waiter wants, and the guest correctly “infers” that the waiter wants the menu, rather than, say, the napkin or a tip. This well-practiced action may originally have been paired with the question “May I take your menu, ma’am?” (which considers the guest’s knowledge and preferences), but over time this utterance and its attendant inferences became superfluous, as myriad interactions led to the desired result. The reaching action has acquired a powerful role in the whole script, triggering the guest’s desired response and making mental state inference unnecessary—unless there is some doubt. If the menu lies on the far side of the guest’s place setting, the script will have to be adjusted, and the waiter may reintroduce some sort of question, because he believes the guest might not know what he wants. Such adjustments can, in the end, become part of an alternate script, again relieving the agent of any (explicit or perhaps even implicit) mental state inferences.

Let me apply these three sources of preemptive information (self, generic knowledge, scripts) to a fascinating example offered by Bavelas and Coates (1992). Two strangers, involved as participants in an experiment, are told by the experimenter to give an opinion on some topic. After the experimenter leaves the room, A says, “You go ahead.” Before A’s last word is completed, B already smiles. A immediately laughs, B says, “Gee, thanks,” and A responds with “You’re welcome.” Within a couple of seconds, the two have conducted a sophisticated conversation in multiple channels and correctly decoded ironic meaning. In particular, B’s inference that A is being ironic is almost instantaneous, and so is A’s recognition that B understands the irony. It seems likely that B didn’t want to “give her opinion” and (implicitly) assumed that A didn’t either; she also may have quickly searched for a way to get out of being the first to speak. So when A said “You go ahead” (perhaps with slightly exaggerated generosity), B could use her own reluctance as the basis of inferring

A's reluctance and therefore code A's utterance as a not-so-generous offer and the specific formulation as displaying generosity with irony. That is, the context and B's own feelings allowed her to interpret A's move as one of copping out, but the specific words he used were of a different script (the "generous offer") and were therefore understood as ironic. The rest of the conversation, with the irony mutually known, is played out according to the cultural script of responding to a generous offer ("Thanks"—"You're welcome"), with smiles, laughs, tone, and added words (e.g., "Gee . . . ") confirming the continued irony.

If facile mindreading occurs between strangers, we shouldn't be surprised that longtime couples can perform apparent feats of mindreading—such as one partner completing the other's sentence or answering a question before the other has even posed it. Long-term relationships benefit from improvements in all three sources of information. Through shared experiences and converging preferences, one's own mind becomes a more reliable indicator of the other's mind; one gains not only generic but agent-specific knowledge; and there are literally thousands of scripts that are practiced in the relationship every day. Long-term relationships also remove a notorious limitation of interactions among strangers: One learns in which contexts the other person does *not* feel or want the same thing as oneself does or will *not* act as other people do.

So, how can we reconcile the apparent low level of many mental states inferences with the seeming "high-level" character that many prototypical mental state inferences show? Are the processes that subserve this rapid system the same as the ones that engage conscious, deliberate reasoning about others' minds? I believe that the only way we can account for the full range of mindreading is by postulating not one mechanism that comes in degrees of conscious awareness, but a whole set of psychological tools that serve mindreading functions (see Ames, Chapter 10, this volume; Fernandez-Duque & Baird, Chapter 5, this volume). Some are fast and general, others are slow but aim at precision even in new situations. Some rely on stored knowledge of trends and patterns; others rely on the perceiver's own mental states in the specific context.

CONCLUSION: MINDREADING AS A MANIFOLD

All three puzzles considered here suggest that the processes underlying mindreading form a *manifold,* a complex array of related but distinct elements. In light of the first puzzle (the role of behavior in mental state inferences), I argued that there were multiple entries into the noncircular inferential relationships between behavior and mental states: the intentionality concept, transparent behaviors, and "effect" states. In light of the second puzzle (how to talk about the mind), I suggested that there are multiple criteria for appropriate mental state ascriptions (introspection, observation, memory, logic, etc.). None of these criteria "defines" the meaning of mental state terms, but the convergence and social negotiation of the evidence at hand determines the successful application of such terms. In light of the third puzzle, finally, I concluded that people rely on multiple tools when dealing with other minds—tools that include explicit mental state inferences but also more implicit processes, such as emotional contagion, behavioral mimicry, and assumptions in conversation.

The picture that emerges is one of mindreading as a diverse toolbox that covers a broad range of stimuli, information processing mechanisms, and outputs. This toolbox includes a conceptual framework (e.g., the intentionality concept and distinctions among representational states as well as emotions); behavior observation capacities (e.g., for eye gaze, basic actions, and emotional expressions); and the ability to imitate and synchronize behaviors, emotions, and physiology. Add to that capacities I haven't discussed here—joint attention and joint action, imagination and pretense, and explicit perspective taking. The list could easily be expanded, and in many cases we don't yet know the fundamental processes or mechanisms that support the specific functions. But it seems clear that no simple notion of a "mindreading module" or an all-encompassing "theory" of mind will do the job of accounting for what people do when they make sense of other minds. As a social species, humans have evolved a large number of paths to other minds that provide both redundancy and flexibility to achieve their interaction goals in many different contexts and under many different demands. Even though the words we use to describe this fascinating phenomenon (*mindreading, mental state inference, theory of mind*) suggest a singular, bounded process or ability, only the recognition of manifolds at all levels—functional, cognitive, and neurological—will help us understand this unique and wondrous characteristic of human nature.

REFERENCES

Astington, J. W. (1993). *The child's discovery of the mind*. Cambridge, MA: Harvard University Press.

Baron-Cohen, S. (1999). The evolution of a theory of mind. In M. C. Corballis & S. E. G. Lea (Eds.), *The descent of mind: Psychological perspectives on hominid evolution* (pp. 261–277). New York: Oxford University Press.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. (2000). *Understanding other* *minds: Perspectives from developmental cognitive neuroscience* (2nd ed.). New York: Oxford University Press.

Barsade, S. G. (2002). The ripple effects: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, *47*, 644–675.

Bavelas, J. B., & Coates, L. (1992). How do we account for the mindfulness of face-to-face dialogue? *Communication Monographs*, *59*, 301–305.

Bavelas, J. B., Black, A., Lemery, C. R., & Mullett, J. (1986). "I show how you feel": Motor mimicry as a communicative act. *Journal of Personality and Social Psychology, 50*, 322–329.

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher–student interactions. *Journal of Nonverbal Behavior, 12*, 120–138.

Blackburn, J., Gottschewski, K., George, E., & L—, N. (2000, May). *A discussion about theory of mind: From an autistic perspective*. From Autism Europe's 6th International Congress, Glasgow. Retrieved April 10, 2003, from www.autistics.org/library/AE2000–ToM.html

Bogdan, R. J. (2000). *Minding minds: Evolving a reflexive mind by interpreting others*. Cambridge, MA: MIT Press.

Buss, A. R. (1978). Causes and reasons in attribution theory: A conceptual critique. *Journal of Personality and Social Psychology*, *36*, 1311–1321.

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen- through 18–month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development, 21,* 315–330.

Carver, C. S., Ganellen, R. J., Froming, W. J., & Chambers, W. (1983). Modeling: An analysis in terms of category accessibility. *Journal of Experimental Social Psychology, 19*, 403–421.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*, 893–910.

Chartrand, T. L., Maddux, W. W., & Lakin, J. L. (2005). Beyond the perception–behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. In R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 334–361). New York: Oxford University Press.

Friedman, H. S., & Riggio, R. E. (1981). Effect of individual differences in nonverbal expressiveness on transmission of emotion. *Journal of Nonverbal Behavior, 6*, 96–104.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, *62*, 378–391.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*, 165–193.

Givón, T. (2005). *Context as other minds: The pragmatics of culture, sociality and communication*. Amsterdam: Benjamins.

Goldman, A. I. (2001). Desire, intention, and the simulation theory. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 207–225). Cambridge, MA: MIT Press.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion*. New York: Cambridge University Press.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, *17*, 187–245.

Jeannerod, M., & Frak, V. (1999). Mental imaging of motor activity in humans. *Current Opinion in Neurobiology*, *9*, 735–739.

Levenson, R. W., & Ruef, A. M. (1992). Empathy: A physiological substrate. *Journal of Personality and Social Psychology*, *63*, 234–246.

Levenson, R. W., & Ruef, A. M. (1997). Physiological aspects of emotional knowledge and rapport. In W. Ickes (Ed.), *Empathic accuracy* (pp. 44–72). New York: Guilford Press.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3*, 21–43.

Malle, B. F. (2002). The relation between language and theory of mind in development and evolution. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 265–284). Amsterdam: Benjamins.

Malle, B. F., & Knobe, J. (1997a). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33,* 101–121.

Malle, B. F., & Knobe, J. (1997b). Which behaviors do people explain? A basic actor–observer asymmetry. *Journal of Personality and Social Psychology, 72,* 288–304.

Malle, B. F., Moses, L. J., & Baldwin, D. A. (Eds.). (2001). *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: MIT Press.

Malle, B. F., & Pearce, G. E. (2001). Attention to behavioral events during social interaction: Two actor–observer gaps and three attempts to close them. *Journal of Personality and Social Psychology, 81,* 278–294.

McClure, J. (2002). Goal-based explanations of actions and outcomes. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 12, pp. 201–235). New York: Wiley.

Meltzoff, A. N. (2002). Elements of a developmental theory of imitation. In A. N. Meltzoff and W. Prinz (Eds.), *The imitative mind: Development, evolution, and brain bases* (pp. 19–41). New York: Cambridge University Press.

Meltzoff, A. N., & Brooks, R. (2001). "Like me" as a building block for understanding other minds: Bodily acts, attention, and intention. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 171–191). Cambridge, MA: MIT Press.

Patten, S. B. (1999). Epidemics of violence. *Medical Hypotheses, 53*, 217–220.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Povinelli, D. M. (2001). On the possibilities of detecting intentions prior to understanding them. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 225–248). Cambridge, MA: MIT Press.

Read, S. J. (1987). Constructing causal scenarios: A knowledge structure approach to causal reasoning. *Journal of Personality and Social Psychology*, *52*, 288–302.

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

Russell, J. (1996). *Agency: Its role in mental development*. Hove, UK: Erlbaum.

Ryle, G. (1949). *The concept of mind*. London: Hutchinson.

Sabini, J., Garvey, B., & Hall, A. (2001) Shame and embarrassment revisited. *Personality and Social Psychology Bulletin, 27*, 104–117.

Sabini, J., & Silver. M. (2005).Why emotion names and experiences don't neatly pair. *Psychological Inquiry, 16,* 1–10.

Silver, M., & Sabini, J. (1978). The perception of envy: An experimental analysis. *Social Psychology, 41,* 105–118.

Simner, M. L. (1971). Newborn's response to the cry of another infant. *Developmental Psychology*, *5*, 136–150.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wellman, H. M., Hickling, A. K., & Schult, C. A. (1997). Young children's psychological, physical, and biological explanations. In H. M. Wellman & K. Inagaki (Eds.), *The emergence of core domains of thought: Children's reasoning about physical, psychological, and biological phenomena* (pp. 7–25). San Francisco: Jossey-Bass.

Whiten, A. (1999). The evolution of deep social mind in humans. In M. C. Corballis & S. E. G. Lea (Eds.), *The descent of mind: Psychological perspectives on hominid evolution* (pp. 173–193). New York: Oxford University Press.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Malden, MA: Blackwell.

Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, *22*, 145–160.

3

A "Constituent" Approach to the Study of Perspective Taking

What Are Its Fundamental Elements?

MARK H. DAVIS

One of the pleasures of participating in an interdisciplinary volume such as this is the opportunity to take a step back from focusing on the kinds of specific questions that are most easily amenable to empirical investigation and to examine instead a favorite topic in a broader, more comprehensive fashion. In this chapter I take such a look at the topic of perspective taking. My hope is that doing so will provide new insights into the way that social psychologists have studied this phenomenon—and how we might do so more effectively.

My chief argument in this chapter is that there are certain essential components of any perspective-taking effort and that it is valuable to consider how these components have been empirically addressed. While I readily admit that there are other ways to parse the phenomenon, it seems reasonable and useful to identify four core constituents of any perspective-taking attempt. These constituents emerge in four questions that can be asked about any attempt by a person to imagine the internal state of another:

- What is the purpose, or *aim*, of the perspective-taking attempt?
- What *sources of information* are used in the perspective-taking attempt?

- What are the *processes* employed during the perspective-taking attempt?
- What are the *results* of the perspective-taking attempt?

Figure 3.1 provides a visual representation of these constituents, arranged roughly in chronological order.

WHAT IS THE AIM OF THE PERSPECTIVE-TAKING ATTEMPT?

It seems likely that most of our attempts to imagine the perspective of another person are directed toward some specific goal. That is, perspective taking is not usually a random affair but instead has a reasonably precise aim. Frequently this aim is to determine another person's *thoughts* or *emotions* (see Figure 3.1). For example, trying to figure out what a spouse is thinking during an ominously silent meal is one common example of such an aim. Sometimes the aim may be to understand the target's *perceptual* point of view—how the world literally looks to him or her. Perspective taking may also be carried out to infer others' *goals, motives,* or *intentions*. Why, for example, does your friend seem to care so much about something that you consider a trivial matter?

I should note at this point that the aims identified in Figure 3.1 could be organized in other ways. Goals and motives, for example,

Aim	**Information Used**	**Process Employed**	**Result**
Thoughts Emotions Perceptual Point of View Motives Goals Intentions	*Target* Face and Body Vocal Cues Words and Actions *Environment* Physical Features Social Features *Observer* Prior Experience Cognitive and Emotional States Knowledge about Social Categories	Mimicry Associative Processes Projection Logical Inference Simulation Imagination	*Cognitive* Accuracy Attribution Self–Other Merging *Emotional* Sympathy Distress Emotional Match *Motivational* Forgiveness Valuing Other *Behavioral* Helping Less Aggression

FIGURE 3.1. A proposed model of the constituents of perspective taking.

could be considered a subcategory of thoughts/emotions rather than as distinct categories. I have separated them here to highlight the difference between thoughts and feelings regarding one's *current* situation, and more general motivational states and goals that may transcend situations. However we define them, what all of these aims have in common is that they involve subjective states of the target and thus are largely invisible to the observer. We cannot directly see another's thoughts, intentions, or goals; they have to be inferred in some way from other information available to us. And that brings us to the second constituent of perspective taking.

WHAT SOURCES OF INFORMATION ARE USED IN THE PERSPECTIVE-TAKING ATTEMPT?

In order to achieve our perspective-taking aims—to infer the unseen—we need something visible, or at least apprehensible, with which to work. There are certainly many different kinds of information that we can use, but it may be useful to think of three broad categories. First, in many cases the *target* will provide the richest source of information—facial expressions, posture, voice, statements, actions, and so on. But other sources of information are also available, either instead of or in addition to the data provided by the target. The *environment* in which the target exists—both physical and social—provides a second source of information that can be used in perspective taking. For example, salient environmental features (e.g., a looming deadline at your spouse's workplace) can be highly useful in explaining a target's behavior (e.g., ominous dinnertime silence). Finally, the *observers* themselves can provide important information. An observer's prior experience with the target, prior experience in similar situations, current cognitive and emotional states, and knowledge about social categories to which the target belongs are all sources of information residing within the observer that can be used as part of the perspective-taking effort. Of course, once we have all of this information, we have to do something with it, and that represents the third constituent.

WHAT PROCESSES ARE EMPLOYED DURING PERSPECTIVE TAKING?

This constituent of perspective taking is concerned with the way in which the estimates of another's internal states are actually produced. The processes that have been linked with perspective taking, theoretically and empirically, vary considerably in their cognitive sophistication and in the degree to which they are under the control of the individual. For example, some processes can be thought of as essentially automatic and perhaps even unconscious. The involuntary physical *mimicking* of others’ facial expressions would fall into this category, as would *associative processes* such as the evocation of classically conditioned emotional responses to target cues. Other processes may be to some degree automatic but also somewhat subject to the observer’s control, such as the *projection* of one’s own states or traits onto the target. The most deliberate and controlled processes include such mechanisms as *logical inference, simulation*, and *imaginative processes* in which observers deliberately attempt to construct in their own minds what the experience of the other is like. What all of these processes have in common is that they are phenomena that take place inside the observer, prompted by exposure to another, and that ultimately lead to some kind of outcome.

WHAT ARE THE RESULTS OF PERSPECTIVE TAKING?

Thus, the final constituent of perspective taking is its outcome: the *result* of the perspective-taking effort. Many different outcomes have been predicted to result from perspective taking, and a number of these have been supported empirically. Some of these outcomes are clearly cognitive in nature: greater accuracy in judging the target (e.g., Bernstein & Davis, 1982), changes in explanations offered for the target’s behavior (e.g., Regan & Totten, 1975), the creation of observer–target merging (e.g., Davis, Conklin, Smith, & Luce, 1996), and reduced activation or application of stereotypes (e.g., Galinsky & Moskowitz, 2000). Some outcomes of perspective taking are emotional: greater feelings of sympathy for a distressed target (e.g., Toi & Batson 1982), greater feelings of personal anxiety or distress (e.g., Schaller & Cialdini, 1988), and greater sharing of emotional states between observer and target (e.g., Miller, 1987). Other outcomes are motivational: a greater readiness to forgive another for a transgression (e.g., McCullough, Worthington, & Rachal, 1997) or a greater valuing of the target’s welfare (Batson, Turk, Shaw, & Klein, 1995). Finally, some outcomes of perspective taking are behavioral: increased helping (e.g., Underwood & Moore, 1982), decreased aggression (e.g., Richardson, Green, & Lago, 1998), or improved social effectiveness as a result of successfully anticipating the other’s reactions (e.g., Davis & Kraus, 1991).

Now, a caveat. This organizational scheme possesses a shortcoming inherent to all such models. It oversimplifies the world and is thus subject to some important limitations and qualifications. To take just one example, in cases where automatic processes like mimicry are engaged, it is not clear that perspective taking can be said to have a conscious *aim* in the way that more deliberate perspective-taking efforts do; the aim in such cases is more implicit, rooted perhaps in our evolutionary history as a highly social species. There are no doubt other qualifications to the model as well. However, this approach does have one important feature: It provides a new vantage point from which to examine the ways that psychologists, especially social psychologists, have studied perspective taking. Just as seeing an aerial view of your own neighborhood causes you to see the familiar in an unfamiliar way, thinking about perspective-taking research from a novel vantage point leads to some insights that would otherwise go unnoticed. For the remainder of this chapter, I will briefly describe three of these.

SOME CONSTITUENTS FORM NATURAL COMBINATIONS

In theory an overwhelming number of different aim–information–process–result permutations are possible; however, careful consideration also suggests that certain combinations are much more likely than others. Although some aims undoubtedly can be attained through the use of multiple information sources, and some kinds of information can be subjected to a variety of different processes, it also seems likely that many of the pathways implied by Figure 3.1 are relatively uncommon and that the number of pathways actually utilized is much lower.

For example, consider the following situation, depicted in Figure 3.2. An observer is attempting to infer the emotional state of a target who is displaying overt facial and bodily cues. It seems quite likely in this case that the aim of *inferring emotion* will lead to a focus on *target* information (including facial cues), which will then prompt automatic processes such as *mimicry*, and that the end result will be *affective changes* in the observer—perhaps greater sharing of the target's emotional state. Indeed, research suggests that a focus on target faces often leads to mimicry (Dimberg, 1990; Vaughn & Lanzetta, 1980) and that such mimicry leads to affective changes in the observer (Adelmann & Zajonc, 1989).

Certainly, attending to target cues can also lead to processes more cognitively sophisticated than mimicry. For example, we often puzzle over the meaning of facial expressions in a very controlled and deliberate way—during a first date, for example, or a poker game. Such efforts rely heavily on logical inference and problem solving rather than mimicry. However, there often seems to be a kind of primacy for automatic pro-

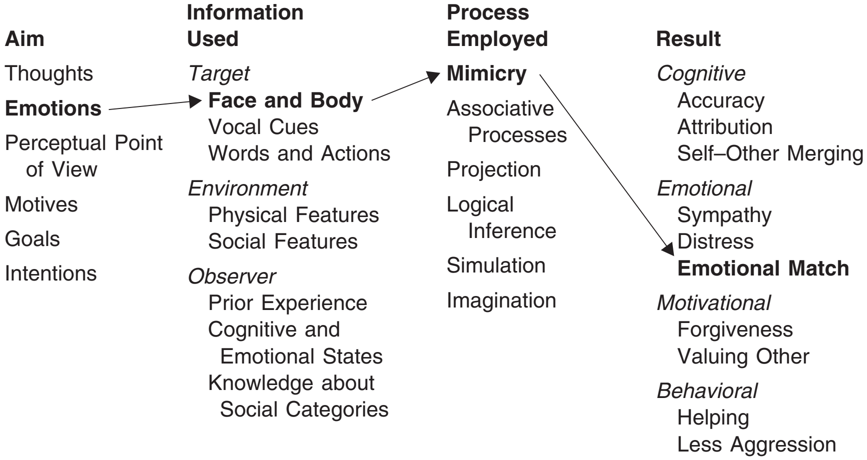

FIGURE 3.2. An example of a "natural" combination of perspective-taking constituents.

cesses when we encounter vivid target information; the presence of powerful social cues evokes automatic mimicry rather quickly, and before other, more cognitively sophisticated, processes. Why would this be? One explanation can be seen in the argument advanced by Preston and de Waal (2000) that the fundamental mechanism underlying empathy is a perception action mechanism (PAM)—a biological tendency, when observing the state of another, to automatically activate one's internal representations of that state, which in turn generate autonomic and somatic responses in the observer. According to Preston and de Waal (and others—see, e.g., Gallese, 2003), there is good evidence that evolutionary forces have supported the development of such a mechanism in humans. If so, then humans may be hard-wired to respond in somatic ways to observed emotions in others.

There are other natural combinations of constituents as well. For example, consider a situation in which the observer's aim is to discern the target's *perceptual point of view*, perhaps within the context of the Three Mountains Task (Piaget & Inhelder, 1956), in which one person is seated in front of a model of mountains and asked to discern how these mountains look to other people seated in different places. Such an aim is likely to lead to a focus on *salient environmental information* (rather than target facial expressions) and to a process such as *mental rotation*, eventually producing a more *accurate estimate* of the target's sensory experience.

In fact, it may be that particular elements in the proposed model of perspective taking are tightly enough linked—logically, or even physiologically—so as to produce characteristic perspective-taking "chains" in which the presence or activation of certain features tends to predictably generate specific processes or attention to particular kinds of information. An exploration of this possibility is one promising avenue for future perspective-taking research.

SOME CONSTITUENTS HAVE RECEIVED INSUFFICIENT ATTENTION

A second fact that becomes apparent when considering perspective taking through the lens of the constituent approach is that empirical investigations have tended to focus attention heavily on certain parts of this perspective-taking model while largely ignoring others. For example, consider the category of "aims." The focus in social psychological research has been almost exclusively on one part of this category: ascertaining targets' current thoughts and feelings. The most common approach in this research tradition has been to induce perspective taking by means of instructional sets that explicitly direct the observers to imagine the target's thoughts, emotions, or both (e.g., Davis et al., 1996; Stotland, 1969). In essence, the instructions in such research *provide* the observers with an aim for their perspective-taking efforts. However, almost never do the instructions direct the observers to imagine anything else—the target's motives or goals, for instance—and that is a bit curious. Why wouldn't aims such as these at least occasionally be the focus of investigation?

In large part, the answer may have to do with methodology. Targets in this sort of research are frequently presented to the observer via videotape or audiotape (e.g., Coke, Batson, & McDavis, 1978; Regan & Totten, 1975); sometimes targets are depicted in written vignettes or scenarios (e.g., Betancourt, 1990). Thus, there is almost never any interaction between observer and target. Obviously, such a procedure provides a huge advantage in terms of experimental control. At the same time, though, there is a cost; the target is very much a "static" entity—not physically present, and not interacting with the observer. Given how little information the observers usually possess about the target in these investigations, it is perhaps not surprising that perspective-taking instructions are directed at what may be considered the most obvious and somewhat superficial of aims—what the target is feeling *right now*, for example—rather than considering more complex social aims such as target intentions and motives. (It is possible, of course, that even in such situations the observers go beyond their instructions and make inferences about goals or motives. Typically, however, the dependent measures employed in these investigations do not allow such inferences to be reported.) It would be interesting to see the result of instructing observers to discern target motives and intentions instead of current cognitive and emotional states.

It is possible to carry out similar analyses for each of the other three constituent categories: information, processes, and results. In each case it is clear that certain features have received sustained research attention and others have not. To take just one more example, consider the "information" constituent. As already noted, a considerable number of studies have explored the ways in which the target's facial expressions influence observers. More recently, observer information has also begun to attract some interest. Davis and colleagues (1996), for example, examined the degree to which perspective takers will attribute self-descriptive traits to novel targets; within the framework of the constituent approach, this is an instance in which at least some of the information used in the perspective-taking attempt (knowledge about self-traits) resides wholly within the observer. In a similar vein, Galinsky and Moskowitz (2000) investigated how perspective taking influences the use of group stereotypes when observers attempt to infer the traits of a target belonging to that group. The researchers found that perspective taking decreased the application of stereotypic information to targets, in large part because of a greater use of self-traits. While the activation (or application) of the observer's knowledge structure can certainly be conceived of as a process, my emphasis here is on the content of the knowledge structure as a source of *information* residing within the observer.

In contrast, the remaining type of information—environmental information—has received much less systematic study, despite the fact that observers almost certainly make use of several kinds of environmental information when attempting to infer another's internal state. For example, *physical features* of the environment, such as temperature, humidity, or the presence of obvious threats, can have clear implications for a target's thoughts and feelings. ("I imagine he must be unhappy, given the desert heat—not to mention the scorpions.") Similarly, *social features* of the environment will often be useful in discerning another's internal states. Is the target alone in the situation, or are there others? Are these others friends or strangers? Same-sex or opposite-sex? Similar to the target or dissimilar? What are the status relationships in the situation? All of these social-environmental features provide valuable clues to the internal state of the target, and all of them are probably used extensively by real-life observers. However, research to date has yet to make any systematic effort to understand the role of such information in perspective-taking attempts.

THE UNINTENDED EFFECTS OF PERSPECTIVE TAKING

The third realization to which the constituent approach gives rise is perhaps the most surprising: the discovery that, to a considerable degree, what social psychologists have studied for the past three decades are really *the unintended effects of perspective taking*. When we explicitly consider the aim of most perspective-taking efforts, it is clear that what is being sought by the observer is usually some kind of accuracy; what the observer desires is some better insight into the internal state of the target—thoughts, or motives, or goals. However, when we consider the other end of this model—what *results from* perspective taking—it turns out that a huge proportion of what social psychology has focused on has nothing to do with this kind of accuracy. Dozens of investigations have been carried out to determine the effect of perspective taking on observers' emotional reactions, attributions, helping behavior, aggression, stereotype use, and so forth. As interesting and important as all of these phenomena are, they are also *not* typically the aim of real-world observers. Only the cognitive result of "accuracy" (see Figure 4.1) seems to reflect the perspective taker's usual goal. In short, although it is not impossible that real-world observers sometimes have aims such as "behaving less aggressively" or "reducing stereotype use" when they engage in perspective taking, it seems much more likely that their usual goal is to gain some accurate insight into the target. If this is true, then it this raises some interesting questions. For starters, why *haven't* social psychologists studied the intended aims of perspective taking?

Of course, one answer to that question is that they have. The question of accuracy has not been completely ignored—it just hasn't always been addressed as consistently as the unintended consequences. For example, research has been conducted over the past 60 years—albeit sporadically—into the question of what makes someone an accurate judge of others (e.g., Taft, 1955). In recent years, moreover, investigators have begun to examine the accuracy question in increasingly complex and sophisticated ways (e.g., Ickes, 1997; Levenson & Ruef, 1992). However, relatively little of this work has focused on the specific role of perspective taking (or its constituents) in producing such accuracy.

The other answer to this question, I believe, is that social psychologists as a group have traditionally been more interested in particular outcomes that might be influenced by perspective taking than we have been in perspective taking itself. So, for example, when attribution theory was dominant, perspective taking's effect on attributions received attention (e.g., Gould & Sigall, 1977; Regan & Totten, 1975); when people have been interested in altruism and helping behavior, perspective taking has been used as a way to induce such actions (e.g., Batson, 1991); as the phenomenon of forgiveness has attracted recent interest, investigators have incorporated perspective taking into their models as one contributing factor (e.g., McCullough et al., 1997). In short, since most empirical interest in perspective taking has tended to be tactical—as a means of examining some *other* phenomenon—surprisingly little attention has been devoted to understanding perspective taking in its own right.

A second intriguing question is this: How frequently do people engage in perspective taking for their own accuracy-oriented reasons but end up with some unintended consequence instead? For example, can an observer take the target's perspective in order to defeat him in a negotiation and unwittingly begin to experience sympathy or an increased sense of oneness with that target? For those with a taste for the ironic, this is a very interesting possibility. In fact, hints of it may be seen in some research by Batson and his colleagues (Batson et al., 1997), in which observers who were led to take the perspective of an unsavory target (e.g., a convicted murderer) came to hold more favorable attitudes—not only toward that particular target, but toward the entire stigmatized class to which he belongs! It seems likely in this case that these more tolerant attitudes toward murderers as a class were an unintended outcome of perspective taking; whether this happens in more natural settings as well remains to be seen, but the possibility seems worthy of examination.

CONCLUSION

This brief consideration of perspective taking through the lens of a constituent approach suggests an interesting conclusion: that the study of perspective taking by social psychologists has been in some ways a haphazard enterprise, and the result has been a rather unsystematic accumulation of knowledge. Thinking about perspective taking in a more organized and formal fashion is a way to combat this problem, and the constituent model presented here is one attempt to do so. There is, however, nothing magical about this particular approach; there are undoubtedly equivalent ways to accomplish the same goal. Rachel Karniol (1986; Karniol & Shomroni, 1999), for example, has offered an approach to perspective taking that is also very process-oriented, although much more narrowly focused than the constituent model presented here. The crucial element in any such approach, however, is to make perspective taking *itself* the focus of more consistent research attention. As long as we study perspective taking only as a means to an end, we may end up never knowing what it truly means.

REFERENCES

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

Batson, C. D. (1991). *The altruism question: toward a social-psychological answer.* Hillsdale, NJ: Erlbaum.

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? *Journal of Personality and Social Psychology, 72*, 105–118.

Batson, C. D., Turk, C. L., Shaw, L. L., & Klein, T. R. (1995). Information function of empathic emotion: Learning that we value the other's welfare. *Journal of Personality and Social Psychology, 68*, 300–313.

Bernstein, W. M., & Davis, M. H. (1982). Perspective-taking, self-consciousness, and accuracy in person perception. *Basic and Applied Social Psychology, 3*, 1–19.

Betancourt, H. (1990). An attribution-empathy model of helping behavior: Behavioral intentions and judgments of help-giving. *Personality and Social Psychology Bulletin, 16*, 573–591.

Coke, J. S., Batson, C. D., & McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of Personality and Social Psychology, 36*, 752–766.

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. *Journal of Personality and Social Psychology, 70*, 713–726.

Davis, M. H., & Kraus, L. A. (1991). Dispositional empathy and social relationships. In W. H. Jones & D. Perlman (Eds.), *Advances in personal relationships* (Vol. 3, pp. 75–115). Greenwich, CT: JAI Press.

Dimberg, U. (1990). Facial electromyography and emotional reactions. *Psychophysiology, 27*, 481–494.

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*, 708–724.

Gallese, V. (2003). The roots of empathy: The shared manifold hypothesis and neural basis of intersubjectivity. *Psychopathology, 36*, 171–180.

Gould, R., & Sigall, H. (1977). The effects of empathy and outcome on attribution: An examination of the divergent-perspectives hypothesis. *Journal of Experimental Social Psychology, 13*, 480–491.

Ickes, W. (Ed.). (1997). *Empathic accuracy.* New York: Guilford Press.

Karniol, R. (1986). What will they think of next? Transformation rules used to predict other people's thoughts and feelings. *Journal of Personality and Social Psychology, 51*, 932–944.

Karniol, R., & Shomroni, D. (1999). What being empathic means: Applying the transformation rule approach to individual differences in predicting the thoughts and feelings of prototypic and nonprototypic others. *European Journal of Social Psychology, 29*, 147–160.

Levenson, R. W., & Ruef, A. M. (1992). Empathy: A physiological substrate. *Journal of Personality and Social Psychology, 63*, 234–246.

McCullough, M. E., Worthington, L. L., Jr., & Rachal, K. C. (1997). Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology, 73*, 321–336.

Miller, R. S. (1987). Empathic embarrassment: Situational and personal determinants of reactions to the embarrassment of another. *Journal of Personality and Social Psychology*, *53*, 1061–1069.

Piaget, J. & Inhelder, B. (1956). *The child's conception of space*. London: Routledge & Kegan Paul.

Preston, S. D., & de Waal, F. B. M. (2001). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences, 25*, 1–72.

Regan, D. T., & Totten, J. (1975). Empathy and attribution: Turning observers into actors. *Journal of Personality and Social Psychology, 32*, 850–856.

Richardson, D. R., Green, L. R., & Lago, T. (1998). The relationship between perspective-taking and nonaggressive responding in the face of an attack. *Journal of Personality, 66,* 235–256.

Schaller, M., & Cialdini, R. B. (1988). The economics of empathic helping: Support for a mood management motive. *Journal of Experimental Social Psychology*, *24*, 163–181.

Stotland, E. (1969). Exploratory investigations of empathy. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 4, pp. 271–314). New York: Academic Press.

Taft, R. (1955). The ability to judge people. *Psychological Bulletin*, *52*, 1–23.

Toi, M., & Batson, C. D. (1982). More evidence that empathy is a source of altruistic motivation. *Journal of Personality and Social Psychology*, *43*, 281–292.

Underwood, B., & Moore, B. (1982). Perspective-taking and altruism. *Psychological Bulletin*, *91*, 143–173.

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. *Journal of Personality and Social Psychology, 38*, 909–923.

4

Starting without Theory

Confronting the Paradox of Conceptual Development

DANIEL D. HUTTO

INTRODUCING THEORY-THEORY

For over two decades the dominant theory, in both philosophical and psychological circles, about how we humans (and disputably apes) get by in our everyday affairs has been that we ascribe thoughts, desires, and intentions to others, drawing on a "theory" or "set of principles" about such mental states. When successful, such ascriptions are putatively the means by which we causally explain what lies behind outward behavior and action.[1] This hypothesis has become known as the *theory*-theory: it is called such because the idea that commonsense psychology has a theoretical (or principled) basis is just one theory about its nature—albeit the dominant one.

There are many versions of theory-theory, which vary over such issues as its scope and the precise way in which the principles in question are represented. Nevertheless, the distinguishing feature of all versions is that the basis of our capacity for understanding others is assumed to be essentially *conceptually* grounded. Accordingly, we are thought to navigate the social world by deploying *principles* of folk psychology and using principles of other kinds for different domains (see Gopnik & Meltzoff, 1997). Defenders of theory-theory are thus committed cognitivists, maintaining that properly *intelligent* activities are best explained by the fact that we—or better, at some level, "our minds"—consult contentful principles, specific to given domains, that guide us in judging and acting. These principles serve as regulative bodies of knowledge.

In setting out what he calls folk psychology's "action principle," Botterill gives a clear example of such a rule. He writes:

> If belief–desire psychology has a central principle, it must link belief, desire and behaviour. It could be formulated like this: An agent will act in such a way as to satisfy, or at least to increase the likelihood of satisfaction, of his/her current strongest desire in the light of his/her beliefs. (1996, p. 115)

Still, one need not go so far as Bottterill in thinking that we must "have some sort of awareness of the principles involved" (Botterill, 1996, p. 114). Minimally all that is required is that our intelligent responses be governed by causally operative, subpersonal, "innately cognized propositional contents" (Fodor, 1983, p. 5).

Proponents of this type of view are unashamedly intellectualist in their commitment to "domain specific propositional attitudes, not [just] . . . domain specific 'devices' " (Fodor, 2001, p. 107). For example, Jerry Fodor has argued that the postulation of conceptually grounded principles that guide our responses is precisely what separates representational from purely mechanistic approaches to psychological explanation. For him, attributing this kind of higher-order "knowledge" to creatures and systems is "the characteristic feature of contemporary cognitivist theorizing" (Fodor, 2001, p. 109). He asserts that, "nobody has the slightest idea how what a creature knows could determine its behaviour unless the propositional knowledge of its knowledge is mentally represented" (Fodor, 2001, p. 114). To be propositional such knowledge must be conceptually grounded because concepts are the constituents of propositions, which are deployed in judgments. To judge (or mentally represent) that "*X* is a Spartan" requires one to have (i.e., possess) the concept *Spartan*, just as to judge (or mentally represent) that "*X* believes that *P*" requires one to have the concept *belief*. Of course, there are differing views about what constitutes the possession of a concept, and taking stock of these will become important later.

I fully agree that unless this distinction is observed one will be hard-pressed to justify the first use of "theory" in theory-theory. It is important to stress this, since a major ambition of this chapter is to set out some reasons for preferring the idea that we start our folk psychological careers with built-in, domain-specific mechanisms rather than an early theory of mind.

THE PARADOX

There is a paradox about how our understanding of one another develops, if we take seriously the cognitivist idea that our intelligent social interactions have robust conceptual underpinnings of the sort described above; for young infants demonstrate a capacity to reliably detect, attend, and respond to others' intentions very early on. Recent experimental evidence confirms that by 10 or 11 months of age they have the capacity to appropriately parse the mentalistic joints of what would otherwise be an undifferentiated behavior stream, showing evidence of distinguishing object directness and action connectedness circa 12 months (Wellman & Phillips, 2001; Woodward, Sommerville, & Guarjardo, 2001). By 14–18 months, they are able to recognize intentions in the sense of being able to effectively distinguish actions that are performed on purpose from mere accidents or movements (Carpenter, Akhtar, & Tomasello, 1998; Meltzoff, 1995). If we follow the cognitivist trend in thinking that to have such a capacity automatically requires having the relevant concepts, then, at first glance, it seems we should ascribe to these infants at least some basic mental concepts and *the* concept of *intention* in particular.

But there is a problem. Sticking with what is required for the attribution of the concept of intention, Astington observes:

> There is something paradoxical about intentional attribution. Though it is true that motivational states are more obvious and more frequently inferred than beliefs, the attribution of intention—in a precise and complete sense—is not simpler, and may indeed be harder, than the attribution of belief. The paradox might lie in the fact that, although children seem to have some command over some aspects of the concept of intention early on, not until later years do they seem to have command over other aspects. (2001, p. 86)

The trouble is that if we hold that the concept of intention is constituted by its links with other mentalistic concepts, and in particular the concept of belief, which children do not fully master until around age 3 or 4, then we ought to conclude that infants *lack* the full-fledged concept of intention at this tender age since they lack the other relevant concepts that constitute its content.

While this is not a problem for every form of cognitivism, it is a serious one for theory-theorists, per se. It is a problem for them precisely because they hold that the concepts that constitute their theoretical principles get their meaning or content holistically, which is to say the content of a given concept is constituted by its links with other concepts. It is worth digressing in order to remind the reader just how central this idea is to theory-theory. Its defenders hold the having of meaning holistically to be the very mark of a concept being *theoretical*. They maintain that the very content of theoretical concepts is determined by the role they play in the network of principles that, working in unison, enable prediction and explanation (and possibly interpretation). David Lewis (1978), one of theory-theory's philosophical fathers, was the first to claim that terms such as *belief*, *desire*, and *hope* are theoretical terms *because* they gain their meaning by virtue of the way they systematically interrelate with one another and relate to other terms that are not part of the theory in question. He illustrated this point with the example of *Cluedo* (or "Clue," as it is called in the United States). The familiar characters in that game—Professor Plum, Colonel Mustard, Miss Scarlet, etc.—can be thought of as the core elements or basic ingredients of a "suspect theory." They are the basis upon which we form our meaningful hypotheses about who is responsible for the death of the victim in the game.

To apply suspect theory we must have some grasp of the rules (or laws) about how these entities can stand in relation to one another and to other entities that are not themselves potential suspects, for this is the background against which we can speculate about the culprits. Thus, we can plot the logical space of possibilities in the Cluedo/Clue world because these entities enter into constraining logical relations with one another and other entities in their universe. We know, for example, that if Professor Plum is in the drawing room he can't be in the study, we know that he cannot have the rope because Colonel Mustard has it, and we know that he can't have the lead pipe since Miss Scarlet has it, and so on. In fact, the entities of this game are fully defined or constituted when these sorts of potential relations to the others are made actual (or become known). Crucially, our understanding of the good Professor Plum is fully exhausted by his role in the game world, and Lewis (1978) and subsequent theory theorists have held that this is equally true of our mental concepts. Concepts like *belief* and *desire* are constituted by their relation to each other and by their characteristic relations to nontheoretical entities and events: their typical causes and effects, as determined by the "laws" of folk psychology. A key difference is that Cluedo/Clue "suspects" all have the same basic roles, but our mental concepts play different *kinds* of roles, making folk psychology a more sophisticated sort of game. Hence, removing one or another of these basic conceptual ingredients, like removing specific chess pieces from play, would not only reduce the number of speculations one might make, it would also dramatically alter the character of those speculations.

The important point is that it is on this basis that our familiar mentalistic vocabulary (i.e., our talk of thoughts, feelings, and expectations) has been understood as being like other theoretically embedded vocabularies (e.g., talk of electrons, atoms, and gravity). Mentalistic concepts are thought to have meaning in precisely the same way as explicitly theoretical terms: holistically. For example, in renouncing the empiricist view that meaning is infallibly supplied by certain types of experience, Paul Churchland famously advanced a general argument that all our judgmental encounters with the world (and one another) are mediated by theories of different kinds:

> If the meaning of our common observation terms is determined not by sensations, but by the network of common beliefs, assumptions, and principles in which they figure, then, barring some (surely insupportable) story about the incorrigibility of such beliefs, assumptions, and principles, our common observation terms are semantically embedded within a framework of sentences that has all the essential properties of a theory. (1979, p. 37; cf. Goldman, 2001, pp. 212–213)[2]

It is their alleged holistic nature that makes folk psychological concepts theoretical and not merely cognitivist in character. One of the most attractive features of holism is that it affords a basis for maintaining that our concepts and categories are not unalterably fixed. Conceptual schemes develop and change over time. Our categories concerning "what there is" are plastic, pliable, and mutable. Indeed, it is precisely because our conceptions of the world shift and change that we can make the sort of rare conceptual advances that constitute the progressive march of science and the growth of knowledge. Based on this kind of view, arch theory-theorists Gopnik and Meltzoff have long maintained that we initially have a basic starter theory of mind in place and that it develops over the course of childhood in exactly the same way that scientific theories develop (Gopnik & Meltzoff, 1997, p. 26).

But accepting just this brings us face to face with the paradox. Against this background, it can be stated quite clearly; for we must ask: How can a person both have *the* concept *X* at time t_1 but still only acquire *the* concept *X* at time t_2? Strictly speaking, if we stick with theory-theory, to avoid contradiction one must accept, as Astington does, that the person must have two different—if related or perhaps somewhat overlapping—concepts since their content, as fixed by their different surrounding theories, must also be different. Certainly, if the full-fledged concepts such as *belief* and *rational thinker*, for example, are not available until the mature stage, we have established that the conceptual content of the early and later theories cannot be equated with each other.

But to accept this leaves us with an important question about the content of the concepts of putative "early theory." Holism demands that the contents of concepts are bound up with the theories of which they are a part, and theories are conceptually distinguished by the sum of their parts. Hence, if we strictly adhere to holism, it would be misleading at best to say that children have the concept of *intention* early on at all, other than the fact that some of their capacities reveal at least a capacity to make certain basic distinctions appropriate to having such a concept. Although tempting, we could not even say, more softly—as Astington proposes—that they have only mastered *the* concept of intention to some extent without also breaking faith with holism since concepts cannot be individuated without involving the whole theory of which they are a part. Perhaps it might be speculated that the way out of this problem is to drop the use of the definite article *the*. For example, we might say that, strictly speaking, the child and the adult have concepts of intention with different contours but that these still somehow both fall under the same general class: they are both concepts of intention. But what permits the identification of these two concepts as being of the same kind? It cannot be by appeal to their inferential relations, globally understood, since those are precisely what drive us to differentiate them.

At this point theory-theorists face a three-pronged dilemma: (1) simply admit that there are two quite distinct concepts in play, as the theories and inferential connections in question are not equivalent (perhaps not even related); or (2) deny the developmental facts; or (3) abandon holism (and theory-theory) altogether.

The first option is not attractive; for if one accepts that the content of these concepts is fixed by their extant theories and that these are, *ex hypothesi*, different enough to warrant the claim that their different relations make the concepts in question different, then it will be difficult even to justify the claim that the early and late versions are related without calling on features that take us beyond the inferential confines of the theory, understood globally (for example, by observing that they both target the same psychological phenomenon, namely, intentions). Certainly, until this is addressed, theory-theorists ought to refrain from identifying the two concepts with one another. But this is not satisfactory since we want to place these two seemingly distinct concepts along a single developmental trajectory. Hence, what we really want is an account of what holds constant about them in the roles they play in both the early and late theories: What about them stays the same throughout the change in their surroundings? It is precisely their relationship to each other over time that we are most interested in. Note, for these purposes, it is of no use to simply observe that at the early stage children somehow "have" the relevant concept potentially, for if the surrounding core concepts play a constitutive role in determining content, then current concepts cannot be thought of as having their content fixed by merely potential relations to other concepts (even if these are likely to obtain in the future).

I will consider the other two ways of getting around the paradox, setting out the consequences of facing each of the horns in turn.

CHALLENGING THE DEVELOPMENTAL FACTS

It is well known that children get better at wielding their folk psychology as they get older, and this is often taken as evidence that their concepts are developing. A major stage in this process comes between the ages of 3 and 5, during which time children normally gain a full understanding of propositional belief, as evidenced by their grasp of the possibility that beliefs can be false. But it is at least possible that children in fact have the full theory from the start, even though they do not learn how to deploy it fully until they are older. Drawing this kind of competence/performance distinction, one might attempt to get around the paradox by maintaining that young children have only a partial grasp of the very same folk psychological principles and concepts that adults employ in later life. Fodor advances this sort of line, favoring the idea that our mature folk psychological theory is in place from the beginning, such that "the child's theory of mind undergoes no alteration; what changes is only his *ability to exploit what he knows*" (Fodor, 1995, p. 110). Although the whole theory is already there, young children use only some of the theory's basic principles. For example, they might use only heuristic H1 and not H2 (set out below) even though both are present in their theory of mind:

H1 Predict that an agent will act in a way that will satisfy his desires.
H2 Predict that an agent will act in a way that will satisfy his desires,
 if his beliefs were true. (Fodor, 1995, p. 112)

Or, more in keeping with theory theory, it might be supposed that young children are initially using *different* precursor principles and concepts than the ones that will eventually "unfold" in the mature phase. Thus, as Segal has argued, our theory of mind might evolve in this way over time through a series of staged diachronic modular upgrades. He sees conceptual development as a preset nonrational process akin to the growth of hair or teeth in certain species; the explanation for these developments is "built into our genes. . . . The modularity theory thus reduces conceptual development in childhood to a kind of process that is reasonably well understood in general terms, and applies across a very wide range of phenomena" (Segal, 1996, p. 155).

These accounts are both nativist, although they challenge the developmental facts in different ways. The first supposes that the mature theory already exists in its full form, long before children have complete facility in using adult psychological concepts in predictions or explanations. The second supposes that the mature theory is somehow in place "potentially" such that it will roll out over time in a predetermined way. Therefore, if either of these positions is to be rendered fully plausible, we are owed an account of how the core framework of principles became established *prior* to our engaging in sophisticated folk psychology. For most naturalists, the answer is that we inherit our fully developed theory of mind (though perhaps not an immediate capacity to use it all or immediate full possession of all its components) from our evolutionary forefathers.

Baron-Cohen goes so far as to claim that the development of an innate "mindreading" module was the ancient solution to an adaptive problem that arose for our ancestors during the Pleistocene epoch (Baron-Cohen, 1995, p. 12). And Segal proposes that

> It is not implausible to suppose that the early theories of mind—the two-year-old's, three-year-old's and so on—are *a hangover from phylogenetically prior stages*. . . . The false representations of, say, the three-year-old are a product of a *relatively primitive system that was present in the species some while ago*. The system was, indeed, good enough for survival. (1996, p. 156, emphases added)

But there is a high price to be paid for adopting either of these accounts, since it is not good enough to say only that the *early* theory was a hangover from our ancient ancestors. If we are to appeal to our evolutionary endowment to explain diachronic stages of "theory change" that converge on a common end point, then our ancestors must have had the *full-fledged* theory. This is not plausible for a number of reasons that I have set out elsewhere (Hutto, 2006).

REJECTING HOLISM ABOUT CONTENT

In addressing the same problem in another context, Nichols and Stich (2003) have simply abandoned the idea that conceptual content is determined holistically. They write:

> When our ancestors first had the capacity to deploy the Goal and Strategy strategy did they really have the concept of *goal*? Our answer here is a familiar one: Yes, all the critters had something like the modern con-

> cept. But since important components of that are part of the modern process of mindreading were not yet in place, their concepts and ours had rather different "causal powers" or "conceptual roles." (p. 68)

They go on to ask rhetorically, "So at what point do their concepts and ours count as really being the concept of *goal, desire, belief*, etc.? On our view, there is no principled way of answering this question, and it is not a question we propose to take seriously" (Nichols & Stich, 2003, p. 68). Still, as we have seen, it is a serious question deserving of attention. A clear consequence of their proposal is that the concepts in question cannot be determined solely by their conceptual roles. The alternative is that some substantive concepts have at least part of their content, if not all of it atomistically—which is to say, the contents are determined not by relation to other contents but by what the concept is meant to pick out, denote, or stand for. Since this can be fixed independently of how the concept relates to any others, to adopt such a view would explain how a concept could be the one and the same concept even if it finds itself in quite different theories. If concepts can be possessed entirely independently of one another, then core folk psychological concepts might be *had* even without being connected to any other relevant concepts at all.

Fodor, who has been the most steadfast defender of conceptual atomism, is utterly open about his allegiances in adopting this sort of view. Qua Cartesian and qua Humean, he tells us: "Having a concept is having something in one's head that serves to represent the objects of one's thoughts. . . . What Cartesians deny is just that our putting our concepts to the uses that we do is constitutive of the concepts or our having them" (Fodor, 2003, p. 21; see also Margolis, 1998). Fodor does not deny that we make use of representations and concepts, but he holds that they have their meaning *first* and find their uses later. He allows that "acquiring the concept CAT does, of course, *require learning* what property is proprietary to cats as such; namely, being cats" (Fodor, 2001, p. 130, emphasis added). But this is disingenuous, for he has a master argument that demonstrates that some concepts are not learnt, since otherwise we would be faced with an infinite regress. I will not rehearse his famous argument for this conclusion here, but suffice to say it drives him to hold that when we learn concepts we must do so by relying upon a preexisting representational medium that already contains them.

But if the concepts that make up this preexisting representational medium are not learned, how do we come by them? For Fodor, even if concepts cannot be learnt (strictly speaking) by a rational psychological process, they can be acquired by a brute-casual or triggering process. Accordingly, we begin life with only protoconcepts that gain their representational content proper by "locking on" to the class of things that they are to be about; they somehow "latch on" to the right extension class in much the way that certain types of young animals are set to imprint on their mothers. For example, according to one recent proposal, once a lock to a dominant class (of, say, *X*s) is established, the internal representations come to stand for all and only *X*s in all future uses (making future representational error possible) (see Prinz, 2002, pp. 249–251; for more detail, see Hutto, 2006).

It should be clear that to accept this is to accept a purely denotational theory of conceptual content (i.e., one in which the content of a concept is nothing more than what it stands for). And Fodor is happy to do so, openly proclaiming that "there is nothing at all to meaning except denotation" (Fodor, 1990, p. 162). Thus, we are brought full circle, as "[Denotational] theories are, on the face of them, *atomistic* about content" (Fodor, 1994, p. 6).

So far, so good. This seems to be the right way to deal with the paradox, since it allows us to say that, despite differences in the inferences they are inclined to draw, both children and adults nevertheless operate with concepts of intention—and that these are related not because they play exactly the same kind of inferential roles within their extant theories but because they are both *about* or *directed toward* the same kind of phenomenon, the having of intentions. The question is: What are the consequences of accepting this kind of atomism about folk psychological concepts?

The first thing to note is that if folk psychological concepts are atomistic they will lack the features that would make them theory-like, since theoretical concepts depend on their place in an entire network for their meanings. Hence, in learning a *theory* one must learn the *rules* for the use of its concepts; for it is clear that if concepts that are to be counted as "theoretical," in any interesting sense, it will require the mastery of a whole nest of inferences involving rules for the use of a range of interrelated concepts. In this respect, it is only if learning concepts is holistic can it be seen as a "rational" process. Hence, to accept atomism about the content of concepts flies in the face of theory-theory and undercuts a key motivation for believing in it.

But this is not all; for, if the story about content acquisition is just one involving brute causal processes, it is ultimately, and perhaps ironically, "adamantly non-cognitivist—to have a concept is . . . simply to have a mental symbol that stands in the appropriate locked relation to a given extension" (Antony, 2001, p. 211). So, at this point, ought not the atomist abandon both theory-theory and cognitivism too? Fodor's response to a line of criticism advanced by Fiona Cowie is importantly revealing on this issue. Cowie (1999) has complained of Fodor's reliance on protoconcepts that there is nothing interesting for these alleged "concepts in waiting" to be. They cannot be concepts, since concepts are individuated by their contents (i.e., semantically) and protoconcepts do not have any contents, so at best they must be understood as some kind of internal mechanisms for concept acquisition. Yet, as Cowie states, "everyone . . . can agree we have innate protoconcepts in this weak sense" (p. 84). At best "protoconcept" appears to be a forward-looking honorific title, but one that is apt to mislead us.

It seems that one is driven to a blatant noncognitivism about the basis for concept acquisition, which is something Fodor would like to resist. By way of reply, he states that protoconcepts are "unactivated innate concepts" that gain their contents when properly triggered. He then supports this idea further, observing:

> On all standard ethological accounts of triggering, part of what is innate in a triggered concept is a *specification of its proprietary trigger.* Since the trigger of an innate concept is both propriety and innately specified, such concepts can be unvacuously *triggered by reference to what would trigger them*. . . . [Thus] the content of protoconcepts is no particular problem for a semantic externalist, so long as he assumes that it supervenes on (possibly unactualised) dispositions. (2001, p. 138, emphases added)

We might well wonder how "the proprietary trigger" is innately "specified" in a way that does not already require preexistence of the very concepts that are meant to be acquired. Here Fodor's own master argument returns to haunt him; for, unless a credible noncircular account of protoconceptual specification can be supplied without invoking representational contents at yet another level down, the account threatens to collapse into purely a noncognitivist, mechanical story about our basic capacities for concept acquisition. Given some of his self-imposed commitments, there are reasons to think that Fodor will ultimately be unable to avoid this conclusion (see, e.g., Mendola, 2003).

A better story to tell about how these mechanisms come to have the capacity to reliably lock onto to their appropriate targets (when conditions are historically normal) would be an evolutionary one rather than a purely causal-informational one. This would be the easiest way to explain how such mechanisms have come to be appropriately responsive to the "right" kinds of triggers. We could at this point attempt to use this sort of account to make sense of full-blown representations, but there is also a less ambitious option.

We could simply surrender the attempt to give an account of the semantics behind our basic responses while nevertheless appealing to some noncognitivist story to explain their *directedness* (see Hutto, 1999, Chs. 3 and 4). That is to say, in the folk psychological case we could adopt a nonconceptualist explanation of why it is that we are naturally responsive to certain mentalistic features, an explanation that retains the required atomistic features. If we give up thinking of these capacities for recognition and response as conceptually based, it becomes easier to account for them by appeal to mechanisms honed by natural selection. This could explain why infants have come to have reliable abilities to parse the behavior stream appropriately without our having to impute any conceptual abilities to them at all at their early stage of development. Their capacities for detecting and recognizing intentional features of the world need not be conceptual at all. They need not even be protoconceptual; rather, they might provide a straightforwardly *nonconceptual* and *nontheoretical* platform for the later development of full-fledged mentalistic concepts.[3] I maintain that if we take this line the paradox of development dissipates. The price is that we must accept that we do not start our careers with a folk psychological "theory" of any kind already in place.

EPILOGUE: CONCEPTS AND CONCEPTIONS

To make this proposal properly attractive, a final word needs to be said about how it, as opposed to its competitors, might better accommodate the fact that conceptual development occurs along a stable trajectory. Ideally, what is needed is an account of how a concept could both change and yet remain the same. The question is: How is it that a set of nonconceptual capacities can provide a sound basis for the eventual development of certain types of concepts? But, despite postulating only nonconceptual underpinnings, the sort of account I advocate is compatible with a tendency to talk in terms of concepts, once these are understood in a particular way.

What it is to "grasp" or to "have" a concept is murkily and ambiguously understood, as these metaphors of active handling and passive ownership imply. Crucially, it is often unclear when we speak of concepts whether or not the focus is on what the concept is about or the psychological means by which we are related to it. Jackendoff reminds us of the source of this problem:

> There is a fundamental tension in the ordinary language term *concept*. On the one hand, it is something in the world: "the Newtonian concept of mass" is something that is spoken of as though it exists independently of who actually knows or grasps it. Likewise, "grasping a concept"

> evokes comparison to grasping a physical object, except that one somehow does it with one's mind instead of one's hand. On the other hand, a concept is spoken of as an entity within one's head, a private entity . . . that can be conveyed to others only by means of language, gestures, drawing, or some other imperfect means of communication. (2000, p. 305)

As we have already seen, Fodor explicates "having" a concept by deploying the model of "having something in your head." Once you "have a concept" you can do things with it, but otherwise not. Thus, he presses the question: "How is one to suppose that a mind can tell a C from a D unless it is already able to think about Cs and Ds?" (Fodor, 2003, p. 19). As before, he holds that we cannot explain what having a concept is by appeal to specific abilities to "sort" and "infer" anything, for if those abilities are to do the required work—if they are of the *right kind*—then they would already be conceptual abilities of the very sort we are trying to explain. For him, to attempt any such ability-based explanation is thus to talk in circles.

But clearly the right response is that at least in some cases telling or reliably distinguishing a C from a D does not require having a prior capacity to *conceptually* classify C's and D's or to have full-fledged thoughts at all. Those cases in which such capacities are deemed to be *conceptual* are distinguished precisely because the activities in question have characteristic features that demonstrate or reveal that one is explicitly drawing appropriate inferences and following through on logical implications and entailments. In the genuinely conceptual case, thoughts are not prior to classificatory activity; rather, they are evidenced by it.

However, it is abundantly clear that Fodor's line of argument only succeeds if we assume that "concepts are basically classifiers, their extensions being determined by dispositions to apply them" (Millikan, 2000, pp. 42–43). This fuels his need to postulate the prior existence of the appropriate concepts that would determine the character of such dispositions, making them—for him—properly recognitional. But we need not adopt this assumption.

It is surely the case that preverbal creatures are designed to be directed at specific things and have the capacity to reidentify things without being able to classify or categorize them. This is surely the case, for example, with the infamous fly-chasing frog, whose capacities for detecting and reidentifying its prey do not involve conceptually based categorical judgments of the sort "That is a fly" (see Hutto, 1999, Chs. 3 and 4). This is, in effect, a kind of basic know-how, an ability to tap the inductive potential of certain relatively stable, reidentifiable environmental kinds. Hence a creature that can reidentify its particular type of prey or its mates will develop many useful but nevertheless fallible expectations about these and their hidden properties, as gleaned from previous encounters with tokens of their kinds. Accordingly, abilities to reidentify substances turn out to be fundamental, and the activities of *classification* and *categorization* are late-developing and parasitic (see Millikan, 2000, p. 34).

In promoting this view, Millikan asserts: "If a concept is genuinely a substance concept, if its extension is really a substance, *this extension is not determined by one's fallible dispositions to recognize it. . . . The extension* of a substance concept *is determined* not by one's disposition (rightly or wrongly) to recognize it but, first, *by the real extent of the substance in nature*" (Millikan, 2000, pp. 33–34, emphases added). In this passage, Millikan is clearly concerned with concepts as they "exist independently" of us in that she is clear that what a concept is about is not fixed by the ways we have of classifying it. Seen in this light there are no grounds for insisting that there must be a representative medium (e.g., concepts in our heads) of precisely the same power as the concepts we are seeking to acquire, learn, or think about. On this view one can think *about X*, in the sense of being able to reliably respond to it without already "having" a concept of it.

This approach is refreshingly different in allowing that having the capacity for reidentification does not require classification or categorization. Hence, we need not presuppose that the creatures in question already "have" or are capable of making use of proprietary concepts during the learning process. Although being able to reidentify such things requires latching onto objective substances and recurring kinds (even mentalistic ones), this can even be achieved by purely nonconceptual means or by using various different kinds of "inferential" means.

That thoughts and words can have the end of targeting the same substance but can do so by different, more or less reliable means makes it easier to deal with otherwise puzzling cases. Hence, it is possible for a child to have a very different way of recognizing and responding to "the very same concept" without this being inferentially mediated or without using the same inferences than an adult might. To avoid confusion we might, following Millikan, call those inferentially mediated ways of reidentifying substances' "conceptions," observing that there can be many different conceptions of one and the same concept. If we are to avoid the mistakes of the theory-theorists, we must however realize that no particular way of identifying substances—no particular conception, that is—"defines the extension of the concept" (Millikan, 2000, p. 11). Since this provides a way of seeing how concepts can stay the same while conceptions change, we are now in a position to dispose of the "paradox of conceptual development." The price to pay is that it goes against the idea that we start our folk psychological careers with a "theory" of mind in hand. Moreover, it opens up the possibility that in some cases the means by which we grasp what a concept is about is not inferentially mediated at all, theoretically or otherwise. In such cases, it would be possible for a creature to latch onto something that has a potential conceptual specification without its having an inferentially mediated conception of it at all.

NOTES

1. The "theory of mind" terminology allegedly made its first appearance in psychology with Premack and Woodward in 1978, and it found its way into philosophy independently in 1980, as introduced by Adam Morton in his *Frames of Mind* (1980), under the guise of theory-theory. However, the core idea can be in Sellars's (1956/1997) influential masterwork *Empiricism and the Philosophy of Mind*.
2. Of course, even if we accept that the content of our folk psychological concepts is constituted holistically, all this shows is that they are similar to theoretical terms in sharing this characteristic. Theoretical concepts could be a subset of the kind of concepts that gain their meaning holistically; hence, inferring that a set of concepts as gaining their meaning from being part of an interrelated network does not automatically license the conclusion that they are theoretical per se, let alone that they are theory-*like* in other respects.
3. This would be completely in line with the view that "action parsing abilities, if possessed by young infants, might critically subserve the ontogeny of genuine intentional understanding" (Baird & Baldwin, 2001, p. 199).

REFERENCES

Antony, L. M. (2001). Empty heads. *Mind and Language, 16*(2), 193–214.

Astington, J. (2001). The paradox of intention: Assessing children's metarepresentational understanding. In B. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 85–104). Cambridge, MA: MIT Press.

Baird, J. A., & Baldwin, D. A. (2001). Making sense of human behaviour: Action parsing and intentional inference. In B. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 193–206). Cambridge, MA: MIT Press.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Botterill, G. (1996). Folk psychology and theoretical status. In P. C. P. Smith (Ed.), *Theories of theories of mind* (pp. 105–118). Cambridge, UK: Cambridge University Press.

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen through eighteen month old infants differentially imitate intentional and accidental actions. *Infant Behaviour and Development, 21*, 315–330.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge, UK: Cambridge University Press.

Cowie, F. (1999). *What's within: Nativism reconsidered*. Oxford, UK: Oxford University Press.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1990). *A theory of content and other essays*. Cambridge, MA: MIT Press.

Fodor, J. A. (1994). *The elm and the expert: Mentalese and its semantics*. Cambridge, MA: MIT Press.

Fodor, J. A. (1995). A theory of the child's theory of mind. In M. Davies & T. Stone (Eds.), *Mental simulation* (pp. 109–122). Oxford, UK: Blackwell.

Fodor, J. A. (2001). Doing without what's within. *Mind, 110*(437), 99–148.

Fodor, J. A. (2003). *Hume variations*. Oxford, UK: Oxford University Press.

Goldman, A. (2001). Desire, intention, and the simulation theory. In B. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality* (pp. 207–224). Cambridge, MA: MIT Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Hutto, D. D. (1999). *The presence of mind*. Amsterdam: John Benjamins.

Hutto, D. D. (2006). *Folk psychological narratives*. Cambridge, MA: MIT Press.

Jackendoff, R. (2000). What is a concept, that a person may grasp it? In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (pp. 305–333). Cambridge, MA: MIT Press.

Lewis, D. (1978). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50*, 249–258.

Margolis, E. (1998). How to acquire a concept. *Mind and Language, 13*(3), 347–369.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 24*, 470–476.

Mendola, J. (2003). A dilemma for asymmetric dependence. *Nous, 37*(2), 232–257.

Millikan, R. G. (1998). A common structure for concepts, individuals, stuffs and real kinds: Mama, more milk and more mouse. *Behavioural and Brain Sciences, 21*(1), 55–160.

Millikan, R. G. (2000). *On clear and confused ideas*. Cambridge, UK: Cambridge University Press.

Morton, A. (1980). *Frames of mind: Constraints on the common sense conception of the mental*. Oxford, UK: Oxford University Press.

Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretence, self-awareness and understanding of other minds*. Oxford, UK: Oxford University Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioural and Brain Sciences*, 1(4), 515–526.

Prinz, J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, MA: MIT Press.

Segal, G. (1996). The modularity of theory of mind. In P. Carruthers & P. Smith (Eds.), *Theories of theories of mind* (pp. 141–158). Cambridge, UK: Cambridge University Press.

Sellars, W. (1997). *Empiricism and the philosophy of mind*. Cambridge, MA: Harvard University Press. (Original work published 1956)

Wellman, H., & Phillips, A. (2001). Developing intentional understandings. In B. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality* (pp. 125–148). Cambridge, MA: MIT Press.

Woodward, A. L., Sommerville, J. A., & Guarjardo, J. J. (2001). How infants make sense of intentional action. In B. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality* (pp. 149–170). Cambridge, MA: MIT Press.

PART II

Reading Behavior, Reading Minds

This page intentionally left blank

5

Is There a "Social Brain"?

Lessons from Eye-Gaze Following, Joint Attention, and Autism

DIEGO FERNANDEZ-DUQUE
JODIE A. BAIRD

The aim of this volume is to explore what it means to understand other minds. Drawing on philosophical, developmental, and psychological perspectives, the chapters in this book address myriad issues, including how an understanding of minds develops, its significance for social interaction, and its relationship to other social and cognitive achievements. Our chapter brings yet another perspective to bear on these issues. Taking a neuroscientific approach, our discussion centers on how understanding other minds—a central aspect of social-information processing—is represented in the brain.

Currently in the literature there are two very different models of how the brain is organized for processing social information. One model posits the existence of a "social brain." In this view, the human mind has evolved a set of domain-specific solutions to particular social problems. According to this view, there exists a set of mental systems for perceiving and reasoning about social stimuli that act independently from the mental systems involved in perceiving and reasoning about nonsocial things. It is the content of the information—its social nature—that determines the dichotomy, rather than the structure of the problem (Cosmides & Tooby, 1992). These social modules are thought to be unique in their input (i.e., the information they process) as well as in their mechanism (i.e., the rules that govern their processing). Finally, their content specificity is ostensibly caused by phylogenetic evolution rather than by familiarity, expertise, or ontogenetic development. In its neuroscientific aspect, the view posits the existence of dedicated brain areas for the processing of social information (Duchaine, Cosmides, & Tooby, 2001).

A different model, with origins in information-processing research, views the mind as a set of subsystems, each of which processes a collection of basic computations. Although the model allows for the existence of domain-specific processes and algorithms, this is not its defining feature. Rather, what dictates which brain areas become engaged are the computational operations necessary to solve the problem. Most cognitive neuroscience research in attention, memory, imagery, and executive functions has followed this information-processing approach (Posner & Raichle, 1994).

Tasks of social relevance often engage the same network of brain areas. At first glance, this appears consistent with the idea of a "social brain." However, those brain areas are sometimes driven by nonsocial tasks too, which contradicts the "social brain" argument. A main goal of this chapter is to put forward a synthesis that accommodates these disparate results. In agreement with information-processing models, we argue that there is no "social brain" but rather isolable subsystems that perform basic computations. But how does this account for the fact that social tasks tend to activate the same network of brain areas? We argue that the recruitment of similar areas by social stimuli happens because social stimuli disproportionately tap into certain basic computations, such as those related to affective processes. Importantly, this correlation between social stimuli and basic computations is not perfect, leaving room for nonsocial stimuli to engage those same brain areas, and also allowing social stimuli *not* to engage those areas on occasion. The absence of a perfect correlation between social stimuli and basic computational operations also makes the distinction between these two levels useful and theoretically important rather than a mere redescription of the same phenomena.

As evidence for our main thesis, we focus on one social skill in particular—namely, the ability to detect and follow eye gaze. We review behavioral and neurological findings on eye-gaze following and discuss its relation to other perceptual skills of social importance, such as facial emotion recognition and face identification. Next, we discuss whether eye-gaze following is related to mental state attributions and joint attention. We end with a discussion of how the impairment of these abilities in autism may both shed light on their typical development and provide answers to how the brain is organized in its processing of social information.

EYE-GAZE PERCEPTION: BEHAVIORAL EVIDENCE

Both in humans and other primates, the ability to follow the direction of gaze is of great importance for social interaction. Primates such as chimpanzees and macaques spontaneously follow the eye gaze of con-specifics, and direction of gaze conveys social dominance (Tomasello, Call, & Hare, 1998). Humans automatically cue their attention in the direction of gaze: 3-month-old infants can follow perceived gaze (Hood, Willen, & Driver, 1998), and even newborns can discriminate between direct and averted gaze, thus suggesting a strong inborn determinism (Farroni, Csibra, Simion, & Johnson, 2002). What is less clear is whether and to what extent these processes are related to mental state attributions. In nonhuman primates, eye-gaze following appears to be a stimulus-driven, nonmentalistic process (Povinelli, Bering, & Giambrone, 2000). In human adults, mental state inferences may not be necessary, even if mental attributions often coexist with gaze following. This counterintuitive fact is well known to any basketball player who has found him- or herself following the adversary's gaze despite knowing the latter's intention to deceive. In experimental paradigms, subjects automatically follow the direction of the eyes even when eye gaze predicts that the target will occur in the opposite location (Driver et al., 1999). These findings, together with the early development of eye-gaze cueing in infancy, suggest a rigid, nonmentalistic system of gaze following that is driven by the physical properties of the stimulus, such as the relation of the dark iris to the white sclera (Ricciardelli, Baylis, & Driver, 2000) and possibly also by the orientation of the head (Langton, Watt, & Bruce, 2000).

EYE-GAZE PERCEPTION: NEUROLOGICAL EVIDENCE

The system for eye-gaze detection is ideal for exploring neural mechanisms of relevance to social interactions even if, in itself, eye-gaze detection is an elementary part of social behavior. For one, eye-gaze perception can be explored with invasive methods in nonhuman primates, yielding highly specific and localized information. Second, since eye-gaze perception is relatively independent from language, it is less problematic to make generalizations across species. Finally, since eye-gaze detection is driven by external stimuli, it is possible to systematically manipulate the system.

The first question to address is whether there is an isolable neural system for eye-gaze detection in its most rudimentary form. Over the last two decades, neuroscience has provided several good examples of that type of direct mapping. Here we briefly mention two such mappings before describing the eye-gaze system. The first example is a region of the posterior temporal lobe, area MT, which has been implicated as a direct neural correlate of perceived visual movement. Neurons in area MT respond selectively to the direction and velocity of movement, and their electrical stimulation biases the behavioral response to moving stimuli (Salzman, Murasugi, Britten, & Newsome, 1992). Area MT is activated by the aware perception of motion (Tootell et al., 1996). The second example is the fusiform face area (FFA), an area of the ventral temporo-occipital cortex that responds selectively to faces. Lesion to the FFA leads to impaired recognition of facial identity (Damasio, Damasio, & Van Hoesen, 1982). Subjective experience of faces activates this area even in the absence of sensory stimuli, as in the case of visual imagery (O'Craven & Kanwisher, 2000).

Along the same lines, several studies have pointed to an area of the posterior part of the superior temporal sulcus (STS) as critical for the encoding of eye gaze. In neuroimaging studies, this area is activated by faces changing the direction of gaze and even by static displays of faces with averted eyes (Allison, Puce, & McCarthy, 2000; Wicker, Perret, Baron-Cohen, & Decety, 2003). Single-neuron recording studies in monkeys reveal some neurons in the anterior part of the STS that respond specifically to gaze direction (Perrett, Hietanen, Oram, & Benson, 1992), and gaze perception is disrupted by lesions to this area (Campbell, Heywood, Cowey, Regard, & Landis, 1990).

EYE-GAZE PROCESSING: SOCIAL MODULE OR INFORMATION PROCESSING?

The literature reviewed thus far illustrates the critical role of the STS in eye-gaze perception. Given the importance of eye-gaze perception in primate social behavior, some may argue that the STS constitutes evidence for the modularity of the social brain. The strongest version of such a modularity claim would argue that the computation of eye-gaze direction is "encapsulated" in the STS. According to this view, other stimulus properties such as facial emotion, and the context in which they are displayed, should have no influence on the system. A less stringent version of the "social brain" hypothesis allows for modulation by other social input (e.g., facial emotion), although not by general resources (e.g., attention). Neither version allows for the processing of social and nonsocial stimuli by common brain areas.

In what follows, we present evidence against both versions of the "social modularity in the brain" hypothesis. Against the strong version of modularity, behavioral and neurophysiological data reveal rich interactions between eye gaze and other facial processes. Against the weaker version of the "social brain" hypothesis, data reveal modulation by general resources and anatomical commonality in the neural instantiation of social and nonsocial processes.

THE RELATION OF EYE-GAZE PERCEPTION TO OTHER ASPECTS OF FACIAL INFORMATION PROCESSING: BEHAVIORAL AND ELECTROPHYSIOLOGICAL EVIDENCE

Even though eye-gaze perception depends critically on the STS, the impact of eye gaze on behavior is modulated by other factors. For example, faces expressing emotion modulate gaze-evoked shifts of attention. Thus, when the person whose eyes serve as a cue looks happy, attention is sustained, but when the person looks angry, attention is relocated somewhere else (Fenske, Frischen, & Tipper, 2004). Furthermore, fearful faces are more effective at cueing attention than neutral faces, at least in subjects who are anxiety-prone (Mathews, Fox, Yiend, & Calder, 2003). Eye gaze also interacts with facial recognition in that faces with direct gaze are easier to identify than faces with averted gaze (Macrae, Hood, Milne, Rowe, & Mason, 2002).

These interactions among gaze, facial emotion, and facial identity also extend to their neural substrates. The activation of the amygdala, a region involved in emotion recognition, is modulated by the direction of gaze. When faces with neutral expression are displayed, direct gaze leads to stronger activation of the fusiform face area than averted gaze and also increases the functional connectivity between amygdala and the fusiform face area (George, Driver, & Dolan, 2001). A likely interpretation is that the emotional valence conveyed by a direct gaze gets encoded in the amygdala, which in turn modulates the fusiform face area via its feedback projections to improve facial identity recognition. The amygdala also appears to be an integral part of the system for eye-gaze detection, as it is activated by eye gaze and its lesion impairs gaze perception (Kawashima et al., 1999; Young et al., 1995).

The amygdala and STS have bilateral anatomical projections, and both areas send projections directly to the orbitofrontal cortex (OFC), an area critical for emotion regulation (Rolls, 1999). Finally, the STS has functional connectivity with regions important for shifting visuospatial attention to the periphery, such as the intraparietal sulcus, thus being an integral part of the neural circuit for eye-gaze following (George et al., 2001).

The behavioral, functional, and anatomical results described above reveal a network of brain structures, including area MT, FFA, STS, amygdala, and OFC, which act in concert to compute many aspects of facial information that are important for social interactions. Much like the findings from other cognitive systems, these results reveal a division of labor, with each area performing a computation that is smaller than the task as a whole. At the same time, each subsystem is heavily modulated by general resources such as the allocation of attention, thus arguing against any type of strong modularity. For example, when the task requires that subjects pay attention to eye gaze, the activity of the STS is increased, while the activation of the FFA, the area that encodes the structural aspect of faces, remains invariant. As expected, paying attention to face identity leads to the opposite pattern (Hoffman & Haxby, 2000).

SOCIAL AND NONSOCIAL PROCESSES: ANATOMICAL COMMONALITY

The literature reviewed thus far argues that the processing of facial information is instantiated by a network of brain structures that includes, among others, area MT, FFA, STS, amygdala, and OFC. These areas are heavily interdependent and modulated by general resources such as attention.

The available evidence also argues against a brain area—or even a network of areas—dedicated to the exclusive processing of social stimuli. For example, area MT responds to biological motion, but it also responds to random dot movement (Salzman et al., 1992). The STS is sensitive to eye gaze, but it is also sensitive to aspects of language (Martin, 2003), semantic knowledge (Chao, Haxby, & Martin, 1999), visuospatial attention, and target detection (Corbetta & Shulman, 2002). The amygdala is important for eye-gaze detection and the recognition of facial expressions of fear, but it also processes many other aspects of fear—for example, Pavlovian classical conditioning and response to aversive tastes and odors (Davis & Whalen, 2001). The orbitofrontal cortex may be important for recognizing certain facial emotions, but it is also critical for modulating nonsocial stimulus–reward associations, such as the relation between key press and food (Blair, 2003; O'Doherty et al., 2003). The role of the OFC appears to be a more general modulation of stimulus–reward associations, of which facial emotions and direct gaze may be special cases. These are just a few examples of how a brain area can be recruited by social stimuli even if the area also participates in other, nonsocial processes. Finally, although parts of the fusiform gyrus are particularly sensitive to faces, which are a type of social stimuli, these same brain areas are also sensitive to nonsocial stimuli that are globally encoded, such as expert recognition of cars or birds (Gauthier & Nelson, 2001).

In sum, there is a network of brain regions disproportionately engaged by social stimuli such as faces. But those areas are also engaged in basic computations of no social relevance. Thus, their engagement by social stimuli is insufficient proof for the existence of a "social brain." Rather, it suggests that social stimuli often carry particular properties—valence, reward value, spatial arrangement of its features—that disproportionately tap certain basic operations.

EYE-GAZE PERCEPTION: A WINDOW INTO MENTAL STATE ATTRIBUTION?

There is no doubt that encoding and following eye gaze are important social skills for both human and nonhuman primates. Nor is there any doubt that normal human adults often make mental state attributions about the eye-gaze patterns they detect: when an agent directs his or her eyes to an object, adults infer that he or she is *seeing* the object (i.e., creating a mental representation that can be used for guiding future behavior, enriching knowledge, and so forth). For normal human adults, the behavioral description (eye gaze) and its mentalistic redescription (seeing) almost always go hand in hand. Moreover, during development, the covariance between the mentalistic and nonmentalistic levels (i.e., between the "looking at" and the alignment of gaze direction and attended object) may play a role in fostering the emergence of mental state attribution. That is, such covariance may help to bootstrap mental state attribution in typically developing human infants who are experience-ready. But there is no principled reason why this should be true in other populations. Newborns, monkeys, and individuals with autism may detect and follow eye gaze without taking the extra step of attributing mental states to such behaviors (Povinelli et al., 2000).

On the whole, knowing about eye-gaze detection tells us little about whether mental state attributions are being made. Similarly, identifying the brain areas involved in eye-gaze detection tells us little about which areas are critical, or even necessary, for the attribution of mental states. Just because behavior and mental states tend to co-occur does not mean that the same brain areas involved in detecting the behavior would also be involved in the attribution of mental states to that behavior. To put it bluntly, computing the direction of gaze is not the same as using the direction of gaze to infer another's intention, nor need it depend on the same brain structures.

DEVELOPMENT: FROM REFLEXIVE EYE CUEING TO JOINT ATTENTION

So far, we have been describing the rudimentary system of eye-gaze detection, but in typically developing humans this rudimentary system becomes part of a more sophisticated system rather quickly in development. At the age of 9 months, infants start using direction of gaze in a flexible manner that takes into account people's communicative intentions, a skill that forms part of what has been labeled "joint attention." Joint attention is the ability to coordinate attention between interactive social partners with respect to objects or events. It reveals an understanding by infants that adults are intentional agents, that is, that adults voluntarily attend to objects and that their attention can be shared, directed, and followed. Besides responding to another's gaze shift, joint attention includes behaviors such as proto-declarative pointing (i.e., pointing to refer to an object, as opposed to pointing to request an object), imitative learning (i.e., acting on objects in the way that adults are acting on them), and social referencing (i.e., infants' reference to adults for information about the approachability, desirability, and other features of objects) (Tomasello & Rakoczy, 2003). All four of these skills—the flexible use of eye-gaze following, proto-declarative pointing, imitative learning, and social referencing—are correlated, develop in synchrony between the ages of 9 and 24 months, and show little influence from environmental variables. Across these behaviors, there seems to be the expression of a single underlying skill—the understanding of persons as intentional agents—which is almost completely absent in other primates (Povinelli et al., 2000; Tomasello & Rakoczy, 2003). As such, it may be the single most important developmental milestone in the understanding of other minds. At the same time, the features that make it so remarkable, namely its early emergence in development and its exclusivity to humans, are the same that make it inaccessible to most of the methods of cognitive neuroscience. Although adult neuropsychology and neuroimaging provide good approximations and hypotheses, it would be erroneous to simply extrapolate from these models to the case of development (Karmiloff-Smith, 1998). Therefore, researchers have been interested in diseases of development, autism in particular, to provide a more complete picture of how these developing abilities map onto brain structures.

DEVELOPMENTAL DISORDERS: THE CASE OF AUTISM

Autism is a neurodevelopmental disorder with a strong genetic component and a heterogeneous neurological substrate. Abnormalities have been reported in the limbic system (anterior cingulate, amygdala, hippocampus, and orbitofrontal cortex), cerebellum, frontal lobes, superior temporal gyrus, and subcortical structures including the thalamus and the basal ganglia (Lord, Cook, Leventhal, & Amaral, 2000). This broad range of neurological abnormalities is matched at the behavioral level by a broad phenotype that includes motor, linguistic, social, and emotional deficits (Joseph, 1999). Individuals with autism exhibit stereotypic and repetitive motor behavior, and their use of language is both delayed and disrupted. At its core, however, autism is a disorder of social interaction and communication.

Poor eye-gaze following is a specific marker of autism, and one of its earliest signs, evident in children as young as 18 months of age. Some children with autism even fail to use gaze as a cue to locate an object (Leekam & Moore, 2001). This is especially remarkable given that chimpanzees, who are incapable of joint attention, can nevertheless use gaze as an instrumental cue (Povinelli et al., 2000). In one study, school-age children with autism were tested in a naturalistic environment for their ability to use eye gaze as an orienting cue. In the autistic group, children with high mental age performed normally, but those with low mental age were impaired relative to developmentally delayed children matched for mental age. Autistic children of low mental age were capable of following gaze when a target object was observable (a skill that emerges at 6 months of age in typically developing infants) but were impaired when the target was absent (a skill that normally develops at the age of 9 months and that indicates the emergence of joint attention; Leekam & Moore, 2001). Another study tested high-functioning 10-year-olds with autism in a computer-based paradigm and found normal cueing effects (Sweettenham, Condie, Campbell, Milne, & Coleman, 2003). Taken together, these two studies suggest that eye-gaze cueing deficits in autism vary according to the severity of impairment in general intelligence.

In contrast, joint attention deficits in autism occur independently of general intelligence and are an early marker of the disease. Children with autism show deviant patterns of reciprocal gaze behavior with their caregivers as well as deficits in the triadic coordination among themselves, adult, and object (Charman, 2003). Interestingly, joint attention in 3- and 4-year-old children with autism is positively correlated with orbitofrontal function, as measured by tasks that engage this region in normal subjects (Dawson et al., 2002). Orbitofrontal cortex is necessary for adding flexibility to stimulus–reward associations (Fellows & Farah, 2003; Rolls, 1999). An inability to assign stimulus–reward associations and flexibly modify them could be detrimental to the development of joint attention, as joint attention depends on social rewards, such as smiles, that are more variable than nonsocial rewards. Consistent with this hypothesis, autistic infants and toddlers prefer highly contingent, nonvariable feedback, while typically developing children instead prefer variable, imperfect feedback (Gergely & Watson, 1999). The contingency hypothesis illustrates the main problem of arguing for a module devoted exclusively to social stimuli. Even if certain brain regions do process social stimuli preferentially, it does not follow that such preferential processing is due to the "social" nature of the stimuli. In the aforementioned example, the driving force is the variability of the stimulus–reward association, which is typical of social stimuli but can be present in nonsocial stimuli as well.

Individuals with autism also show abnormalities in the processing of facial and emotional stimuli. Studies measuring event-related potentials (ERPs) reveal that children with autism are impaired in the discrimination of novel versus familiar faces, despite their normal discrimination of novel versus familiar objects. Relative to healthy adults, adults with autism have reduced response to facial emotions. This includes abnormal activity in the fusiform gyrus, amygdala, and the STS (Critchley et al., 2000). One possible interpretation of these data is that they provide evidence for a social perception module whose impairment in autism accounts for the social deficits exhibited by this group. Although this analysis captures much of what is wrong in autism, it makes the same mistake previously described in our analysis of stimulus–reward associations. Just because social stimuli such as faces happen to activate the fusiform face area, it does not follow that faces activate this area *because* they are stimuli of social relevance. Nonsocial stimuli such as cars and birds can similarly activate this region, provided that observers are experts in recognizing those objects and encode them globally. The critical factor, therefore, seems to be not the social nature of the stimulus but its holistic encoding. Of course, this is not to say that difficulties in face perception, emotion recognition, and gaze following do not have consequences for mental state inferences. On the contrary, such deficits may put children with autism at a severe disadvantage in their development of mental state understanding. Relative to typically developing children, children with autism have difficulties using the eyes as cues for attributing mental states (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) and in using faces to judge the approachability and trustworthiness of people (Adolphs, Sears, & Piven, 2001). There is also evidence that children with autism have difficulty understanding gaze in mentalistic terms (Baron-Cohen, 1995).

So, is autism best characterized as a deficit in a social module or as a disorder of basic processes that apply to both social and nonsocial domains? Individuals with autism have deficits in joint attention and mental state attributions, but they also have motor deficits, visual attention deficits, and deficits in feedback processing, set switching, executive processing, and other nonsocial abilities (Joseph, 1999). Thus, although autism may present first and foremost as a problem of social cognition, it also manifests itself in nonsocial problems. This dual deficit poses a challenge to both the social module view (why should there be nonsocial deficits?) and the information-processing view (why should the deficit be mostly social?). Our proposal is that the solution to this paradox lies not in the brain itself but in the external stimuli that the brain processes. More specifically, we argue that attributes of social stimuli may correlate with basic information-processing computations. According to our framework, a deficit in the basic computation would affect primarily, but not exclusively, the processing of social information. We have already discussed the examples of variable stimulus–reward associations and holistic visual encoding. In both cases, there is a correlation between the type of stimulus (social/nonsocial) and a computational feature (variable/nonvariable reward; holistic/feature-based encoding). We are not claiming that these two basic computations account for all deficits in autism; rather, we use these examples to illustrate the larger point that general computational deficits may account for what appear to be domain-specific deficits.

Stimulus category (social/nonsocial) may correlate with other basic computations, such as those involved in affective processing. Since social stimuli carry more affective valence than nonsocial stimuli, they should disproportionately tap limbic structures such as the amygdala and the orbitofrontal cortex. Limbic areas are activated by affective stimuli such as faces. These brain areas send feedback projections to regions of the temporal lobe including the fusiform gyrus. During normal development, this affective loop is likely to play a role in modulating the plasticity of the fusiform gyrus, which with experience becomes a dedicated system for face recognition (i.e., the FFA). To put it in psychological terms, faces are attractive stimuli that capture the child's attention, thus becoming the focus of preferred processing and, with the passage of time, a stimulus the child is expert with. But for children with autism, faces do not seem to carry their appropriate valence, and thus these children seem uninterested in faces. Within a social module framework, there is little one could do to rectify this problem. Within an information-processing framework, however, it should be possible to pair facial stimuli with nonsocial rewards that are valued by the autistic child and in this way engage the affective loop. Through extensive training in such a paradigm, it might be possible for children with autism to achieve the type of expertise with facial stimuli that typically developing children gain naturally (Carver & Dawson, 2002).

Also related to the affective dimension, type of stimulus (social/nonsocial) may correlate with orienting effectiveness. For example, while the orienting deficit in autism is most severe for social stimuli, such as faces or being called by name, individuals with autism are also impaired in their orienting to nonsocial stimuli, such as a jack-in-the-box (Dawson, Meltzoff, Osterling, Rinaldi, & Brown, 1998). Interestingly, deficits in joint attention correlate with deficits in orienting toward social stimuli but not with nonsocial orienting. This result is consistent with the idea raised earlier that the covariance of mental and nonmental levels (orienting to a location, finding a social stimulus at that location) may foster the mental state attribution process.

CONCLUSIONS

The mind's redescription of social information into basic information-processing computations may prove to be a powerful tool. We mentioned examples in the context of autism, but it is easy to see how the logic can generalize to typical development. Importantly, such redescription provides the system for processing social stimuli with a flexibility that is central to social interactions and that a social module framework cannot easily account for. Imagine that somebody stares at you with an angry face. Depending on the context, you might feel amused (the person is an actor you are watching in a play), frightened (you are in a shady part of town), or disgusted (the person is the leader of the world superpower justifying his next imperialistic move). In other words, it is easy to imagine that one and the same stimulus will be processed very differently depending on the context (Lange et al., 2003). Such context effects cannot be explained by models positing a strict separation between social and nonsocial information. In contrast, an information-processing account can easily accommodate these contextual effects, as it assumes that general cognitive abilities can influence the processing of information. Reconceptualizing the social brain in terms of basic information processes also allows researchers to ask *how*—in computational terms—social information is processed. Finally, it entrenches social neuroscience within the cognitive neuroscience tradition that has been so fruitful over the past two decades. Just as the brain does not make a strict separation between social and cognitive information, neither should brain researchers.

ACKNOWLEDGMENTS

This research was supported by a postdoctoral fellowship (No. F4866) from the Heart and Stroke Foundation of Ontario to Diego Fernandez-Duque.

REFERENCES

Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of Cognitive Neuroscience, 13*(2), 232–240.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: the role of the STS region. *Trends in Cognitive Science*, *4*(7), 267–278.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241–251.

Blair, R. J. R. (2003). Facial expressions, their communicatory functions and neuro-cognitive substrates. *Philosophical Transactions of the Royal Society of London, Series B, 358*, 561–572.

Campbell, R., Heywood, C. A., Cowey, A., Regard, M., & Landis, T. (1990). Sensitivity to eye gaze in prosopagnosic patients and patients with superior temporal sulcus. *Neuropsychologia, 28*, 1123–1142.

Carver, L. J., & Dawson, G. (2002). Development and neural bases of face recognition in autism. *Molecular Psychiatry, 7*, 18–20.

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience, 2*, 913–919.

Charman, T. (2003). Why is joint attention a pivotal skill in autism? *Philosophical Transactions of the Royal Society of London, Series B, 358*, 315–324.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention. *Nature Reviews Neuroscience, 3*, 215–229.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford University Press.

Critchley, H. D., Daly, E. M., Bullmore, E. T., Williams, S. C. R., Van Amelsvoort, T., Robertson, D. M., et al. (2000). The functional neuroanatomy of social behaviour: Changes in cerebral blood flow when people with autistic disorder process facial expressions. *Brain, 123*, 2203–2212.

Damasio, A. R., Damasio, H., & Van Hoesen, G. W. (1982). Prosopagnosia: Anatomic basis and behavioral mechanisms. *Neurology, 33*, 331–341.

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

Dawson, G., Meltzoff, A. N., Osterling, J., Rinaldi, J., & Brown, E. (1998). Children with autism fail to orient to naturally occurring social stimuli. *Journal of Autism and Developmental Disorders, 28*(6), 479–485.

Dawson, G., Munson, J., Estes, A., Osterling, J., McPartland, J., Toth, K., et al. (2002). Neurocognitive function and joint attention ability in young children with autism spectrum disorder versus developmental delay. *Child Development, 73*(2), 345–358.

Driver, J., Davis, G., Ricciardelli, R., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition, 6*(5), 509–540.

Duchaine, B. C., Cosmides, L., & Tooby, J. (2001). Evolutionary psychology and the brain. *Current Opinion in Neurobiology, 11*, 225–230.

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences, 99*(14), 9602–9605.

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.

Fenske, M. J., Frischen, A., & Tipper, S. P. (2004). Faces expressing emotion modulate gaze-evoked shifts of attention. *Journal of Experimental Psychology: Human Perception and Performance*. Manuscript submitted for publication.

Gauthier, I., & Nelson, C. A. (2001). The development of face expertise. *Current Opinion in Neurobiology, 11*, 219–224.

George, N., Driver, J., & Dolan, R. J. (2001). Seen gaze-direction modulates fusiform activity and its coupling with other brain areas during face processing. *Neuroimage, 13*, 1102–1112.

Gergely, G., & Watson, J. S. (1999). Early socio-emotional development: Contingency perception and the social-biofeedback model. In P. Rochat (Ed.), *Early social cognition: Understanding others in the first months of life* (pp. 101–136). Mahwah, NJ: Erlbaum.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience, 3*(1), 80–84.

Hood, B. M., Willen, J. D., & Driver, J. (1998). Adults' eyes trigger shifts of visual attention in human infants. *Psychological Science, 9*, 131–134.

Joseph, R. M. (1999). Neuropsychological frameworks for understanding autism. *International Review of Psychiatry, 11*(4), 309–325.

Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Ito, K., et al.

(1999). The human amygdala plays an important role in gaze monitoring: A PET study. *Brain, 122,* 779–783.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Science, 2*(10), 389–398.

Lange, K., Williams, L. M., Young, A. W., Bullmore, E. T., Brammer, M. J., Williams, S. C., et al. (2003). Task instructions modulate neural responses to fearful facial expressions. *Biological Psychiatry, 53*(3), 226–232.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Science, 4*(2), 50–59.

Leekam, S. R., & Moore, C. (2001). The development of attention and joint attention in children with autism. In J. A. Burack, T. Charman, N. Yirmiya, & P. R. Zelazo (Eds.), *The development of autism: perspectives from theory and practice* (pp. 105–129). Mahwah, NJ: Erlbaum.

Lord, C., Cook, E. H., Leventhal, B. L., & Amaral, D. G. (2000). Autism spectrum disorder. *Neuron, 28*, 355–363.

Macrae, C. N., Hood, B. M., Milne, A. B., Rowe, A. C., & Mason, M. F. (2002). Are you looking at me? Eye gaze and person perception. *Psychological Science, 13*(5), 460–464.

Martin, R. C. (2003). Language processing: Functional organization and neuroanatomical basis. *Annual Review of Psychology, 54*, 55–89.

Mathews, A., Fox, E., Yiend, J., & Calder, A. (2003). The face of fear: Effects of eye gaze and emotion on visual attention. *Visual Cognition, 10*(7), 823–835.

O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience, 12*, 1013–1023.

O'Doherty, J., Winston, J., Critchley, H., Perret, 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.

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London, Series B, 335*, 23–30.

Posner, M. I., & Raichle, M. E. (1994). *Images of mind*. New York: Scientific American Library.

Povinelli, D. J., Bering, J. M., & Giambrone, S. (2000). Toward a science of other minds: Escaping the argument by analogy. *Cognitive Science, 24*(3), 509–541.

Ricciardelli, P., Baylis, G., & Driver, J. (2000). The positive and negative of human expertise in gaze perception. *Cognition, 77*, B1–B14.

Rolls, E. T. (1999). The functions of the orbitofrontal cortex. *Neurocase, 5*, 301–312.

Salzman, C. D., Murasugi, C. M., Britten, K. H., & Newsome, W. T. (1992). Microstimulation in visual area MT: Effects on direction discrimination performance. *Journal of Neuroscience, 12*(6), 2331–2355.

Sweettenham, J., Condie, S., Campbell, R., Milne, E., & Coleman, M. (2003). Does the perception of moving eyes trigger reflexive visual orienting in autism? *Philosophical Transactions of the Royal Society of London, Series B, 358*, 325–334.

Tomasello, M., Call, J., & Hare, A. (1998). Five primate species follow the gaze of con-specifics. *Animal Behavior, 55*, 1063–1069.

Tomasello, M., & Rakoczy, H. (2003). What makes human cognition unique? From individual to shared to collective intentionality. *Mind and Language, 18*, 121–147.

Tootell, R. B., Reppas, J. B., Dale, A. M., Look, R. B., Sereno, M. I., Malach, R., et al. (1996). Visual motion aftereffects in human cortical area MT reveal by functional magnetic resonance imaging. *Nature, 375*, 139–141.

Wicker, B., Perret, D. I., Baron-Cohen, S., & Decety, J. (2003). Being the target of another's emotion: A PET study. *Neuropsychologia, 41*, 139–146.

Young, A. W., Aggleton, J. P., Hellawell, D. J., Johnson, M., Broks, P., & Hanley, J. R. (1995). Face processing impairments after amygdalotomy. *Brain, 118*, 15–24.

6

Visual Cues as Evidence of Others' Minds in Collaborative Physical Tasks

SUSAN R. FUSSELL
ROBERT E. KRAUT
DARREN GERGLE
LESLIE D. SETLOCK

Consider a surgical team in the midst of an operation. As the operation unfolds, the surgeon requires actions, such as the passing of surgical implements, on the part of the scrub nurse and other assistants. In order to coordinate their behavior, the surgeon and scrub nurse must mutually understand what surgical implement is needed at what time. One way they can coordinate is through language—the surgeon can simply say, for example, "Pass the scalpel." They can also communicate via gestures—the surgeon can point to the desired implement. Interestingly, however, observations of surgical teams have discovered that nurses often *predict,* by watching the surgeon's behaviors and the status of the task, what implements the surgeon needs and can have them ready in advance of any verbal or nonverbal requests (Nardi et al., 1993).

Surgical teams perform one example of a *collaborative physical task,* a task in which two or more individuals jointly perform actions on concrete objects in the three-dimensional world. Such tasks play an important role in many domains, including education, design, industry, and medicine. For example, an expert might guide a worker's performance of emergency aircraft repairs in a remote location, a group of students might collaborate to build a science project, or an emergency room team might combine its efforts to save a patient's life. In each case, the success with which group members can perform collaborative physical tasks depends to a large extent on how well they can infer one another's states of mind.

Observational studies of physical collaboration suggest that people's speech and actions are intricately related to the position and dynamics of objects, other people, and ongoing activities in the environment (e.g., Flor, 1998; Ford, 1999; Goodwin, 1996; Kuzuoka & Shoji, 1994; Tang, 1991). Conversations during collaborative physical tasks typically focus on identifying involved objects, describing actions to be performed on those objects, and confirming that the actions have been performed successfully. During the course of the task, the objects themselves may undergo changes in state as people act upon them (e.g., mechanical pieces of a device may become functional as it undergoes repair) or as the result of outside forces (e.g., a patient might start hemorrhaging).

The performance of collaborative physical tasks requires substantial coordination among participants' actions and talk. In face-to-face settings, much of this coordination is managed through visual information. Visual information plays at least two interrelated roles. First, it helps people maintain up-to-date mental models or *situation awareness* of the state of the task and others' activities (Endsley, 1995). This awareness can help them plan what to say or do next and to coordinate their utterances and actions with those of their partners. Second, visual information can help people communicate about the task, by aiding *conversational grounding*—the development of mutual understanding among conversational participants.

Actions and words are signs of state of mind, and they fluctuate in effectiveness depending on communication media and can substitute for one another in the grounding process. Our research explores relationships among visual information, conversational grounding, and the ability of communicators to understand each other's minds. We also explore the role of this mutual understanding in the collaborative process. Communicators' inferences about each other shape their decisions regarding how to situate, frame, and time their conversation about the task at hand.

In the remainder of this chapter we first describe how people collaborate on physical tasks—how they maintain situational awareness and ground their conversations. Then, we present our theoretical framework for analyzing the role of visual cues in situational awareness and grounding. We next present an overview of our research paradigm and some selected findings regarding how visual cues are used to infer others' minds during collaborative physical tasks. We end with some conclusions about the role of visual information in inferring others' minds and directions for future research.

COLLABORATING ON PHYSICAL TASKS

In order for collaborators to provide useful assistance, they must determine what help is needed, when to provide the help, how to phrase their messages of assistance such that the worker understands them, and whether the message has been understood as intended or additional clarification is needed. That is, assistance must be coordinated not only with the worker's utterances but also with his or her actions and the current state of the task.

Consider the following fragment from a conversation in which a helper is telling a novice worker how to attach a bicycle saddle to its seat post using clamps.

HELPER: Now you want to fit the rails of the seat into that groove.

WORKER: I see. How can it fit into?

HELPER: You might want to unscrew those nuts a bit.

WORKER: Oh—OK.

HELPER: It will give you a little more room.

To have this dialogue, the helper needs to overcome several challenges. One challenge is for the helper to identify what the worker is attending to, in order to determine whether an object is part of the joint focus of attention. The helper's use of the definite article in "the rails" and "the seat" and deictic adjectives in "that groove" and "those nuts" depends on knowing the worker's focus of attention, to be assured that he was referring to the rails, seat, grooves, and nuts the helper was manipulating. A second challenge is to make sure that the worker understands an utterance before continuing the conversation. In this example, the worker verbally indicated understanding with phrases like "I see" or "OK." The helper could also infer understanding because he could see that the worker had indeed started to loosen the nuts. Finally, the helper needs to comply with Gricean norms of conversation, such as informativeness and brevity (Grice, 1975). In this case, he does so by using deictic references (e.g., "that groove") along with pointing.

One way helpers decide how to assist their partners is by maintaining *situation awareness* (Endsley, 1995)—a continually updated cognitive representation of the task and environment. Nardi and colleagues (1993) found that nurses continually monitor surgeons and patients to identify which surgical implement to provide. Likewise, in the bicycle repair task, helpers use their awareness of the state of the bicycle—what repairs have been made thus far, with what level of success—to determine what information to present next. One aspect of situation awareness is the awareness of what a partner is currently doing—what actions he or she is performing, with what tools and parts, and with what success. This information allows helpers to draw inferences about others' minds and to determine whether clarification or expansion of the instructions is required. If, for example, the worker is holding the wrong tool, the helper can interject a comment to correct this (e.g., "No, not that wrench, the larger wrench").

A second way helpers decide how to assist their partners is via conversation. As they present their instructions, helpers monitor workers' responses such as acknowledgments of understanding (e.g., "uh huh," "got it") and questions (e.g., "which wrench?"). Pairs work together to ensure that messages become part of their *common ground*, or shared knowledge and beliefs (Clark, 1996; Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986). The term "grounding" refers to the interactive process by which communicators exchange evidence about what they do or do not understand over the course of a conversation, as they develop common ground (Clark & Brennan, 1991). Clark and Marshall (1981) identified three primary sources for grounding: membership in the same group or population (e.g., Fussell & Krauss, 1992; Isaacs & Clark, 1987), *linguistic co-presence* (the dialogue history), and *physical co-presence* (the shared physical setting). Shared views of objects and people are one important aspect of physical co-presence.

Visual cues provided by others' facial expressions, actions, and jointly observable task objects and environment can facilitate situation awareness and conversational grounding. (e.g., Daly-Jones, Monk, & Watts, 1998). In the operating room setting, nurses use visual cues to predict what surgeons will do next and what implements they will need (Nardi et al., 1993). Similarly, helpers in our bicycle repair task can monitor workers' facial expressions, workers' actions, and changes in the state of the bicycle, and tailor their instructions to ongoing changes in the workers' need for assistance. In the excerpt we gave previously, the helper used his observation that the worker was finished with a step to time his next instructions ("Now . . . "), and he used his view of the bicycle to determine that the nuts needed to be unscrewed. Similarly, because workers can view helpers' actions and hand movements, helpers can use pointing gestures and deictic expressions (e.g., "that one") to refer quickly and efficiently to task objects. In the prior excerpt, the helper used a combination of pointing and a deictic expression, "those nuts," to refer effectively to the nuts in question.

Different types of visual cues vary in their importance for maintaining awareness and grounding conversation. For example, seeing a partner's facial expression gives evidence of his or her understanding but provides no information about the state of the task. A view of objects in the environment, in contrast, provides substantial information about the state of the task and indirect cues to others' understanding. A challenge, both for theoretical development and technology design, is to understand how people use specific types of visual evidence to make specific types of inferences about others' minds. What, for example, are the right visual cues to provide to doctors performing remote tele-surgery? Which cues are best for collaborating on an architectural design task across several physical locations?

We take a decompositional approach to this problem, in which we strive to specify the different types of visual information available to collaborators and identify how these cues influence inferences about others' minds, interpersonal communication, and task performance (Kraut, Fussell, Brennan, & Siegel, 2002). Our approach is illustrated in Table 6.1, in which we consider the types of inferences about others' minds that can be drawn from five sources of visual information—participants' faces, participants' head positions and gaze, participants' bodies and

TABLE 6.1. Types of Inferences That Can Be Drawn about Others' Minds Based on Five Types of Visual Cues

Type of visual evidence	Inferences about others' minds
Participants' faces	Facial expressions and nonverbal behaviors can be used to infer level of comprehension, emotional responses.
Participants' head positions and eye gaze	Eye gaze and head position can be used to establish others' general area of attention and infer intended actions.
Participants' bodies and actions	Body position and actions can be used to establish others' focus of attention; appropriateness of actions can be used to infer comprehension.
Task objects	Changes to task objects can be used to infer what others have done.
Work environment	Traces of others' actions may be present in the environment.

actions, the focal task objects, and the work environment. The table is intended to be illustrative of our approach rather than definitive; future research will be needed to fully specify the rows and columns of this table.

When two people are working side by side, they have all five sources of visual information easily available. If the participants have to collaborate across a distance, they must communicate through some type of telecommunications, substantially limiting the shared visual information. Although technology can be a hindrance to smooth interaction, the features of technology allow us to examine how visual evidence affects situation awareness and conversational grounding in ways that are not possible in face-to-face settings. In the next section, we discuss some of our studies of how people make use of visual cues to infer others' minds in collaborative physical tasks.

EMPIRICAL STUDIES OF VISUAL CUES IN COLLABORATIVE PHYSICAL TASKS

To investigate issues of others' minds in collaborative physical tasks, we use experimental paradigms in which pairs work together to perform a task such as repairing a bicycle, building a large toy robot, or completing an online jigsaw puzzle. In each of these tasks, one participant (the "worker") is responsible for manipulating pieces and tools; the other participant (the "helper") provides guidance but does not actually manipulate pieces or tools. The helper is provided with instructions for the task by way of a manual or diagram of the completed project. The task is thus similar to many learning environments, such as teachers instructing students, remote experts guiding vehicle repairs, or tele-surgery.

Within this basic paradigm, we manipulate features of the technologies that the pairs use as they complete the task and compare communication and performance with these technologies to side-by-side and audio-only control conditions. Dependent measures include task performance time, task errors, responses to survey questions about the success of the collaboration and perceptions of the value of the technologies, conversational coding of session transcripts, and, in some cases, more detailed analysis of the videotapes of the sessions.

To date, we have undertaken approximately a dozen studies of communication during collaborative physical tasks (e.g., Fussell, Kraut, & Siegel, 2000; Fussell, Setlock, & Kraut, 2003; Fussell et al., 2004; Gergle, Kraut, & Fussell, 2004a, 2004b; Gergle, Millen, Kraut, & Fussell, 2004; Kraut, Fussell, & Siegel, 2003; Kraut, Gergle, & Fussell, 2002; Kraut, Miller, & Siegel, 1996). Not surprisingly, these studies have consistently demonstrated that conversational grounding is most efficient and performance is best when pairs work side by side and share full visual co-presence. More interestingly, the visual cues provided by most (but not all) video technologies improve grounding and performance over audio-only connections. From our quantitative and qualitative analyses of interactions across different media conditions, we have found substantial evidence to support our theory that video enhances communication and performance by allowing both helpers and workers to use visual cues to infer each other's states of mind. Communicators rely on visual information to understand their partners' changing situations, needs, and priorities.

Evidence That Helpers Infer Workers' Minds

We first consider how visual cues influence helpers in collaborative physical tasks. Our quantitative analyses and more qualitative evaluations of the videotapes from the sessions lead us to propose that helpers make at least three types of inferences based on visual evidence. First, they make inferences about workers' mind states: for example, how busy they are, what their focus of attention is, and whether they understood the instructions. Second, they make more general inferences about worker's abilities—for example, whether they are adept at mechanical tasks or possess the cognitive skills required to perform successfully. Finally, they make inferences about the state of the task—how far along it is, how smoothly things are progressing, and whether or not any emergency situations have arisen. We find that these inferences shape the content and timing of helpers' instructions and their decisions about when to provide additional information or clarifications.

One example we see repeatedly in our videos is helpers' use of visual evidence to infer workers' comprehension of instructions. When they can see workers' actions, helpers provide verbal feedback on the correctness of those actions (e.g., "that's right"). Figure 6.1 shows verbal acknowledgments of behavior in the puzzle experiment. When they can see the worker, even if it is at a delay, helpers provide ongoing feedback about the correctness of those actions; when they can't see the worker, they provide no feedback at all. This suggests that visual cues from others' actions allow people to make inferences about how well they have understood instructions.

Appropriately timed interruptions also provide evidence that the helper is taking into account the worker's state of mind. In a version of our puzzle studies where the pairs used a text-chat client to converse about the task, we saw several instances of the helper having prepared his or her next message and having it ready to be sent. However, when timing was critical and the helper could visually monitor the work space,

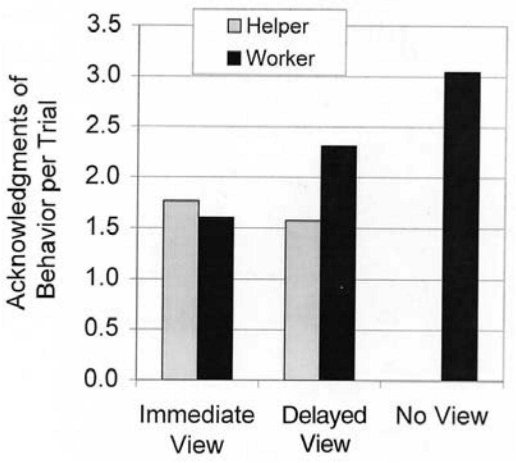

FIGURE 6.1. Acknowledgments of behavior in a collaborative puzzle task as a function of the participant's role and the helper's view of the worker. Data from Gergle et al. (2004).

he or she sometimes waited to send it so that it was received at the appropriate moment (i.e., at a less interruptible time when the partner could attend to the message and when the message matched a particular state of the puzzle).

Evidence That Workers Infer Helpers' Minds

Although the helper's view of the workspace varies with communication medium, the worker's view remains constant across conditions. Thus, if workers failed to attend to helpers' states of mind, they should behave in an identical fashion in all conditions. Instead, we find that workers use the knowledge they have about what helpers can see to infer what helpers know about workers' mind states, abilities, and the state of the task. For example, workers provide different sorts of feedback to helpers depending upon what helpers can see. As shown in Figure 6.2, workers provide significantly more acknowledgments (e.g., "OK," "I got it") when they know that the helper can't see them. When the helper can see the work area, workers are aware that the helper is watching and provide less explicit feedback about comprehension. Similarly, workers use more deictic expressions (e.g., "this one," "here") or pronouns to refer to task objects and locations when they think the helper can see them (side by side or via video) than when the helper is connected by audio only (Figure 6.3).

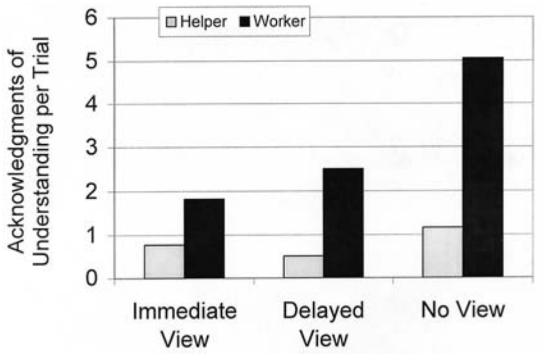

FIGURE 6.2. Acknowledgments of understanding in a collaborative puzzle task as a function of the participant's role and the helper's view of the worker. Data from Gergle et al. (2004).

A compelling example of this behavior can be seen in self-interruptions. In one example, the worker states, "Is it this . . . errr, the one with the black swatch on the right and two green . . . err . . . light green stripes . . . near the top?" In the first part of this utterance, the worker begins to form a deictic reference ("Is it this [one]?") but quickly remembers that the helper cannot see their workspace and replaces it with a more detailed description of the piece (" . . . errr, the one with the black swatch . . . ").

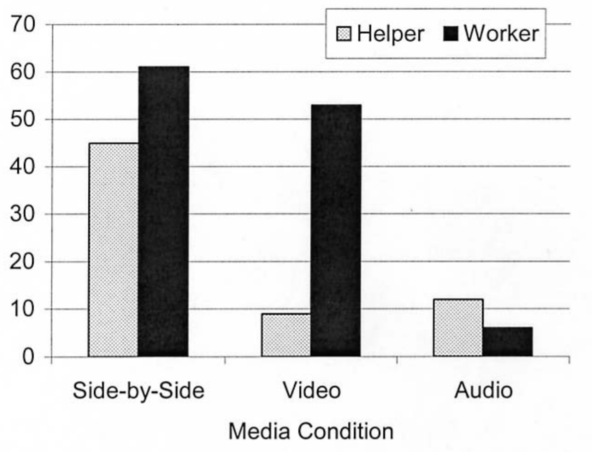

FIGURE 6.3. Use of deixis in a bicycle repair task as a function of media condition. Data from Kraut et al. (2003).

Sources of Evidence about Others' Minds

Results such as those reviewed above clearly demonstrate that visual cues allow communicators to infer each other's knowledge and establish common ground. But which visual cues are most important? In a recent study (Fussell, Setlock, & Parker, 2003), we examined where helpers look as they provide their instructions during a collaborative robot construction task. Helpers wore a helmet with an eye-tracking device during the study. We coded the onset and offset of gaze toward a set of targets: the workers' hands, the workers' faces, the robot being built, the set of task parts and tools, and the instruction manual. We then examined the gaze patterns for regularities that might help us understand how helpers make inferences about workers' mind states.

Our results indicated that helpers looked predominantly at the workers' hands, the robot pieces, and the robot under construction. They looked significantly less at the workers' faces (Figure 6.4). In addition, helpers appeared to use a regular sequence of glances across targets as they provided their instructions. Typically, they first looked at the robot and the manual to determine the next instruction; then, after giving their instructions, they looked at the workers' hands and the robot to

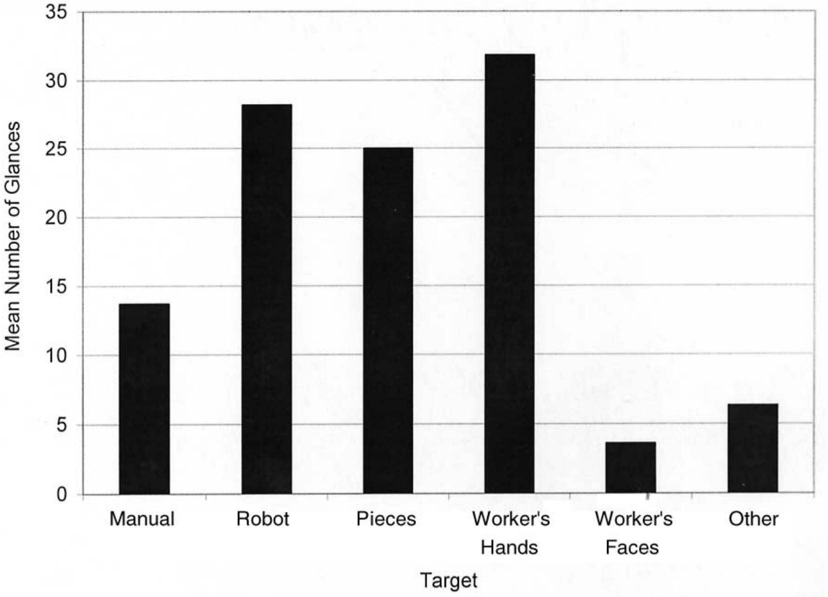

FIGURE 6.4. Mean number of helper glances by target during a collaborative robot construction task. Data from Fussell et al. (2003).

make sure the directions were being followed correctly. The findings support our hypothesis that helpers in collaborative physical tasks draw inferences about workers' mind states by watching what they do and assessing the appropriateness of their actions rather than by observing their facial expressions.

Actions as Evidence of Others' Minds

The results of our gaze tracking study suggest that helpers watch what workers *do* in addition to listening to what they *say* during collaborative physical tasks. The ensuing inferences shape the instructions and feedback they give. In face-to-face settings, much of the worker's contribution to the grounding process may be via his or her actions, as in the following example from our robot construction task:

HELPER: Alright, uh, take the orange thing.

WORKER: (*Picks up the orange dome.*)

HELPER: And, uh, um, one of these little pieces, that one right there.

WORKER: (*Picks up small black piece.*)

HELPER: Put it, uh, on the underside of the orange dome on the left.

WORKER: (*Drops piece, picks it up, connects it to the bottom of the dome.*)

HELPER: Yeah, like that.

We are currently applying sequential modeling to better understand how participants interweave verbal messages and behaviors in the grounding process (e.g., Gergle et al., 2004a). Such analyses permit us to understand in detail how watching others' actions affects conversation and performance. Our results show that when shared visual information is available, helpers use workers' actions in a way that is structurally similar to the way they use verbal statements when no shared visual information is available. For example, in the puzzle task, when a shared view of the workspace was available, the workers were more likely to let their actions "speak" to provide evidence of their comprehension. They were less likely to present verbal acknowledgments both when attempting to select the proper puzzle piece and when positioning a relevant piece within the workspace. The sequential analyses demonstrated that the workers' actions replaced a typical utterance when they knew that the helper could see what they were doing. Similarly, the helpers were more likely to use the workers' actions as evidence of their understanding by simply following the actions with their next description. By using actions as evidence of others' minds, pairs were able to communicate more efficiently.

CONCLUSION AND FUTURE DIRECTIONS

Surgical teams, laboratory partners, and other collaborators on physical tasks must maintain awareness of the state of the task, the status of task objects, and their partners' activities. In addition, they must maintain up-to-date mental representations of their partners' states of mind and current level of understanding. Doing so enables them to make successful inferences about the information needed at any point in time and to ground their messages more efficiently.

One particularly compelling source of evidence about others' minds is visual information. In our work, we demonstrate that helpers use visual cues to infer workers' mind states, workers' general abilities, and the state of the task. Workers, in turn, make inferences about what helpers know based on what they can see. These inferences shape their decisions about how much feedback to provide about their mind states, general abilities, and task status. Both visual information and speech serve as signals of another's state of mind and provide evidence regarding their current level of understanding. This evidence allows the pairs to ground their messages efficiently.

While visual information can signal another's level of understanding, its availability is often affected by the communication media used, which can lead to less efficient communication. While this may be a drawback to everyday communication, it is beneficial from a scientific point of view in that it allows us to understand—at a deeper level—the details of how visual information serves the purpose of successful communication via the modeling of another's state of mind. Our work in this chapter discussed an experimental approach to understanding exactly how the features of the media and task performance interact.

In our current research, we are expanding our model of how visual information is used to infer others' minds in several directions. First, we are examining how properties of joint activities such as the number of participants, sizes and types of objects, and task complexity affect the use of visual information to make inferences about others' minds. Our studies suggest, for example, that visual cues are more important in situations in which objects are difficult to name (Gergle et al., 2004b). The value of specific types of visual information such as views of others' faces or actions is also likely to be task-dependent. Whereas viewing partners' faces was relatively unimportant in the types of tasks we studied, monitoring facial expressions may be much more important for negotiation and other tasks in which assessing others' emotions is essential. Finally, we are examining how cultural differences in reliance on visual cues impacts conversational grounding across different media (e.g., Setlock, Fussell, & Neuwirth, 2004). Our overall goal is to develop a comprehensive model of the role of visual information in assessing others' minds that delineates precisely what visual cues are used, under which task, cultural, and media conditions, to infer specific attributes about others' minds.

In conclusion, visual information provides multiple resources for inferring others' minds during collaborations on physical tasks. Although some of our findings might be attributed to simpler mechanisms, such as mere behavior reading, we believe that such accounts cannot explain the full range of results. In particular, because the worker's field of view is constant across all the media conditions we examined, any differences in workers' behavior must be attributed to their beliefs about what their partners could or could not see. These findings provide strong evidence that theories of others' minds play a crucial role in collaborative tasks.

ACKNOWLEDGMENTS

This research was funded by National Science Foundation Grant Nos. IIS 9980013 and 0208903 and by an IBM PhD Fellowship to Darren Gergle. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies. We thank Susan Brennan, Jane Siegel, Jie Yang, Jiazhi Ou, and Adam Kramer for their thoughtful comments on this work. We also thank our many undergraduate research assistants for their help in conducting studies and preparing data for analysis.

REFERENCES

Clark, H. H. (1996). *Using language.* Cambridge, UK: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, R. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: APA Press.

Clark, H. H., & Marshall, C. E. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge, UK: Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1–39.

Daly-Jones, O., Monk, A., & Watts, L. (1998). Some advantages of video conferencing over high-quality audio conferencing: Fluency and awareness of attentional focus. *International Journal of Human–Computer Studies, 49,* 21–58.

Endsley, M. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*, 32–64.

Flor, N. V. (1998). Side-by-side collaboration: A case study. *International Journal of Human–Computer Studies*, *49*, 201–222.

Ford, C. E. (1999). Collaborative construction of task activity: Coordinating multiple resources in a high school physics lab. *Research on Language and Social Interaction, 32*, 369–408.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, *62*, 378–391.

Fussell, S. R., Kraut, R. E., & Siegel, J. (2000). Coordination of communication: Effects of shared visual context on collaborative work. *Proceedings of the CSCW 2000 Conference on Computer-Supported Cooperative Work* (pp. 21–30). New York: ACM Press.

Fussell, S. R., Setlock, L. D., & Kraut, R. E. (2003). Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. *Proceedings of the CHI 2003 Conference on Human–Computer Interaction* (pp. 512–520). New York: ACM Press.

Fussell, S. R., Setlock, L. D., & Parker, E. M. (2003). Where do helpers look? Gaze targets during collaborative physical tasks. *CHI 03 Extended Abstracts* (pp. 768–769). New York: ACM Press.

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E. M., & Kramer, A. (2004). Gestures over video streams to support remote collaboration on physical tasks. *Human–Computer Interaction, 19,* 273–309.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2004a). Action as language in a shared visual space. *Proceedings of the CSCW 2004 Conference on Computer-Supported Cooperative Work* (pp. 487–496). New York: ACM Press.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2004b). Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of Language and Social Psychology. 23,* 491–517.

Gergle, D., Millen, D., Kraut, R. E., & Fussell, S. R. (2004). Persistence matters: Making the most of chat in tightly-coupled work. *Proceedings of the CHI 2004 Conference on Human–Computer Interaction* (pp. 431–438). New York: ACM Press.

Goodwin, C. (1996). Professional vision. *American Anthropologist, 96*, 606–633.

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (Vol. 3, pp. 41–58). New York: Academic Press.

Isaacs, E., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General, 116,* 26–37.

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human–Computer Interaction, 18,* 13–49.

Kraut, R. E., Fussell, S. R., Brennan, S. E., & Siegel, J. (2002). Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work. In P. Hinds & S. Kiesler (Eds.), *Distributed work* (pp. 137–162). Cambridge, MA: MIT Press.

Kraut, R. E., Gergle, D., & Fussell, S. R. (2002). The use of visual information in shared visual spaces: Informing the development of virtual co-presence. *Proceedings of the CSCW 2002 Conference on Computer-Supported Cooperative Work* (pp. 31–40). New York: ACM Press.

Kraut, R. E., Miller, M. D., & Siegel, J. (1996) Collaboration in performance of physical tasks: Effects on outcomes and communication. *Proceedings of the CSCW 1996 Conference on Computer-Supported Cooperative Work* (pp. 57–66). New York: ACM Press.

Kuzuoka, H., & Shoji, H. (1994). Results of observational studies of spatial workspace collaboration. *Electronics and Communications in Japan, 77*, 58–68.

Nardi, B., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S., & Sclabassi, R. (1993). Turning away from talking heads: The use of video-as-data in neurosurgery. *Proceedings of Interchi 1993* (pp. 327–334). New York: ACM Press.

Setlock, L. S., Fussell, S. R., & Neuwirth, C. (2004). Taking it out of context: Collaborating within and across cultures in face-to-face settings and via instant messaging. *Proceedings of the CSCW 2004 Conference on Computer-Supported Cooperative Work* (pp. 604–613). New York: ACM Press.

Tang, J. C. (1991). Findings from observational studies of collaborative work. *International Journal of Man–Machine Studies, 34,* 143–160.

7

Attributing Motives to Other People

GLENN D. REEDER
DAVID TRAFIMOW

What were President George W. Bush's aims when he ordered U.S. troops to invade Iraq in 2003? What motivated Hilary Clinton to forgive the marital indiscretions of her husband, U.S. President Bill Clinton? As these examples illustrate, questions about the motives of other people arise frequently in daily life. Yet, social psychological research on perceived motives has only recently developed a head of steam (Ames, 2004; Malle, 1999; Read & Miller, 1993; Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002). Perceivers think of motives as mental states that describe the goals and aims of a person's intentional actions. By attributing such motives, perceivers gain some understanding of what a person means in conversation, how the person's actions fit together, and why the behavior occurred in the first place. This chapter outlines some of the emerging issues, with a focus on why, when, and how people infer motives. The first section of the chapter will discuss some of the reasons why people rely on motives to explain other people's behavior. Next, we explore the particular circumstances when people are most likely to infer motives. The last section of the chapter describes some of the psychological processes that underlie inferences about motives. An integrating theme of the chapter is that perceivers often infer motives spontaneously and then use them to make trait attributions about other people.

WHY DO PEOPLE CARE ABOUT MOTIVES?

Dennett (1993) proposed that perceivers have a good reason for approaching others as if they are intentional agents—it works. By treating others as rational beings with common desires and impulses, perceivers are able to predict what others will do. Evolutionary psychologists suggest that early humans who understood the motives of friends and enemies could better anticipate their actions and thereby gained a survival advantage. The ability to recognize enemies and fraudulent schemes may have been particularly important, leading our ancestors to develop a domain-specific "cheater detector" mechanism (Cosmides, 1989). Thus, cave dwellers who recognized that enemies might deliberately draw them away from the cave could take precautions to protect the food stores. According to this line of thinking, modern social perceivers are predisposed to conceptualize others in terms of their goals and motives.

Perceiving a motive enables the perceiver to make sense of behavior patterns that might otherwise appear as random "noise." Heider and Simmel (1944) provided an early demonstration of this sense making by creating a film in which black geometric figures were shown moving against a white background. When a large triangle moved toward a circle in certain ways, perceivers saw the triangle as a "bully" who was chasing the circle. Similarly, by attributing motives to the characters in a book or film, perceivers can organize their understanding of events. For example, to follow the complex plot lines of *The Lord of the Rings*, it is necessary to keep in mind that Frodo, Sam, and Gandolf do not appear motivated by power (e.g., they do not want the ring for themselves), unlike characters such as Sauraman. In sum, inferences about motives allow people to comprehend the meaning of events and better understand how the different aspects of an individual's personality fit together (Read & Miller, 1993).

Finally, we should note that perceivers may prefer motives as explanations because they represent causal powers or generative mechanisms (Johnson, 2001; White, 1995). Thus, we might infer that Harry got a face-lift because he wants to look younger (and increase his appeal to younger women). The motive appears to represent a drive or force that explains the action, similar to the way the force of a strong wind could help to explain why a shingle blew off the roof. Such reasoning contrasts with the type of covariation reasoning proposed by traditional attribution theory (Kelley, 1973). According to covariational logic, a person's behavior tends to be explained by the potential cause with which it uniquely covaries. For example, if getting a face-lift is unique to Harry (other people do not get them), covariation logic suggests the presence of a person cause (e.g., there is something unique about Harry that caused him to behave that way). Johnson (2001) notes that, although the covariation-based approach to causal attribution has dominated the research agenda of social psychologists, perceivers often rely more on generative relations in which effects are produced by causal powers or mechanisms (Ahn, Kalish, Medin, & Gelman, 1995). An explanation in terms of motive can provide just that sort of account. For example, when seeking an explanation for Harry's face-lift, perceivers may find an explanation in terms of motive ("He wants to look young") to be a more specific and satisfying explanation than an explanation in covariation terms ("Something about Harry caused the behavior").

It appears, then, that people have a strong interest in getting to know the motives of others. From an evolutionary perspective, an interest in motives appears necessary for survival in the social environment. In our everyday lives, we need to interpret motives in order to understand the meaning of events and to understand the different facets of a person's personality. As such, inferences about motives offer a particularly satisfying explanation of behavior.

WHEN DO PERCEIVERS ATTRIBUTE MOTIVES AND WITH WHAT EFFECT?

If inferences about motives are both useful and intellectually satisfying, it stands to reason that people should infer them with great frequency. In fact, inferences related to intentionality and motive may proceed more or less automatically. Smith and Miller (1983) asked their research participants to read short sentences (e.g., "Andy slips an extra $50 into his wife's purse") and then answer questions about whether the behavior was intended and whether the behavior was caused by something about the person versus something about the situation. Participants were faster to answer questions about intentionality (2.4 seconds) than questions about either person causality (3.4 seconds) or situational causality (3.8 seconds). Thus, people may be predisposed to process behavior in terms of intentionality rather than in terms of abstract (causal) reasoning about persons and situations (Malle, 1999).

Other studies provide more direct evidence that inferences about motives can be made spontaneously (Fiedler & Schenck, 2001; Kawada, Oettingen, Gollwitzer, & Bargh, 2004; Vonk, 1998). For example, Reeder and his colleagues (Reeder, Vonk, Ronk, Ham, & Lawrence,

2004) presented perceivers with a videotape of a student who was asked by a professor to help move books around in the professor's office. The student readily agreed to the request and began helping with the task. Participants also received information about the situation surrounding the target's helping behavior. In the no-choice condition of the study, perceivers learned that the student was employed by the psychology department to help professors with tasks such as moving books. In other words, moving books appeared to be a normal part of the job. Two additional conditions, however, described the student as responding to different situational forces. In a free-choice condition, the participants were told that the student was not working at the time the help was offered. Finally, in the ulterior motive condition, participants learned that the student had been nominated for a $1,000 academic award, and the professor was the one who would oversee the award. Later, when perceivers wrote out their impressions of the student in an open-ended format, their unprompted descriptions tended to focus on the motives of the target person. For example, participants in the no-choice condition described the student as motivated by (job-related) obedience, whereas those in the ulterior-motive condition described the student as having selfish motives. When forming impressions, then, perceivers may infer motives without an experimenter having to specifically ask about them (see also Reeder, Hesson-McInnis, Krohse, & Scialabba, 2001; Reeder et al., 2002).

A follow-up study examined the speed with which participants made their judgments about motive (Reeder et al., 2004). Once again, participants were provided with information about a student who helped a professor move books. As in the preceding study, the helping behavior took place under no-choice, free-choice, or ulterior-motive conditions. In this study, however, participants received this information as they sat in front of a computer that timed their responses. The results indicated that perceivers were quick to infer a motive that was relevant to the helping situation. For example, those in the no-choice condition were quicker to infer the motive of obedience than were participants in the remaining conditions.

Given that perceivers can infer relevant motives spontaneously, what role do such inferences play in the impression formation process? Additional analyses of the studies reported above suggest that inferences about motive are central to impressions. For example, inferences about motives (such as obedience and selfishness) proved to be good predictors of trait attributions about the student's helpfulness (Reeder et al., 2004). Participants who saw the student as motivated by obedience also tended to think that she was relatively helpful. In contrast, those who saw her as having a selfish ulterior motive rated her as less helpful. In fact, inferences about motive were better predictors of trait attributions regarding the student's helpfulness than were abstract causal attributions (to either personal or situational factors). In sum, perceivers often infer motives spontaneously, and such inferences tend to guide the kinds of trait attributions that are made about the target person.

The findings summarized above are important because, starting with Heider (1958), attribution theorists believed that causal attributions took precedence over other types of inferences. For example, Shaver (1975, pp. 31–33) was quite explicit in specifying a chain of inference in which causal attributions preceded trait inferences. Yet, it now appears that causal attributions (to global person vs. situational factors) are often of secondary importance to perceivers when they explain intentional behavior (Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000) and that inferences about motives play a more prominent role in person perception than earlier attribution theorizing had anticipated (Reeder et al., 2001, 2002, 2004).

Although motives and goals are typically important when perceivers explain intentional behavior, there is a growing body of literature on the specific circumstances under which motive inferences are most commonly made (Idson & Mischel, 2001; Malle et al., 2000; McClure, 2002; O'Laughlin & Malle, 2002; Nussbaum, Trope, & Liberman, 2003). McClure (2002) notes that people are more likely to explain common actions in terms of motive, whereas they tend to explain more-difficult-to-enact behaviors in terms of preconditions (that are thought necessary to enact the behavior). For example, a common behavior such as going to the zoo might be attributed to a motive (such as curiosity about animals), whereas a more difficult behavior such as buying a wildlife preserve would be attributed to preconditions (such as being wealthy). Malle and colleagues (2000) suggest that preconditions can be thought of as including not only enabling factors (such as having wealth) but also explanatory factors that specify the causal history behind the act (e.g., a person grew up surrounded by pets). Other factors, such as familiarity with a target person (as a friend or family member) lead perceivers to increase their use of motive-relevant explanations, as opposed to trait-like explanations (Idson & Mischel, 2001). In addition, Nussbaum and colleagues (2003) suggest that events in the near future (e.g., Will the target person study tonight?) are more likely to be thought of in terms of strategic, motive-relevant constructs, whereas events in the distant future (Will the target person go to college?) are thought to be determined by a person's abstract traits.

In much of the research summarized above, there is an implicit assumption that inferences about motives are an *alternative* to other types of person-relevant explanations. For example, Nussbaum and colleagues

(2003) implicitly assume that perceivers will explain behavior either in terms of concrete mental states such as motives or in terms of abstract traits. In our view, however, perceivers are likely to rely on more than one type of explanation when accounting for a given behavior. The multiple inference model (MIM) of dispositional inference is explicit on this point (Reeder et al., 2002, 2004). MIM suggests that perceivers infer both the motives and traits of a person, integrating the different types of information in a meaningful way. For example, in the two Reeder and colleagues (2004) studies discussed earlier, perceivers inferred a more helpful trait if they attributed the student's helping behavior to an obedience motive, as opposed to a self-serving motive. Thus, rather than viewing motive attribution as an alternative to other types of person-relevant explanations (such as trait attributions), we suggest that perceivers often infer multiple attributes in a target person and then integrate these different types of information.

In summary, recent studies suggest that inferences about motives are not a hothouse phenomenon, observable only when experimenters inquire about them. To the contrary, perceivers often infer motives spontaneously as a way of making sense of behavior. Although traditional attribution theory focused on abstract causal attribution (to situational and dispositional factors) as underlying the traits that we attribute to others, it now appears that inferences about specific motives often play a more important role. Finally, we hope future research will better clarify how inferences about motive are combined with other types of person inferences. Below, we turn our attention to the processes by which motives are inferred.

HOW DO PEOPLE INFER MOTIVES?

Traditionally, philosophers and developmental psychologists (Goldman, 2001; Gopnik & Wellman, 1994) have proposed two process-related explanations to account for inferences about the mental states of others. One explanation suggests that perceivers mentally place themselves in the other person's shoes, attempting to simulate what the target thinks (simulation theory). The other explanation suggests that perceivers invoke general knowledge and implicit theories about other people to infer the target's mental states (implicit theory). In the past, writers have applied these theories to account for mental state inferences regarding the beliefs and feelings of others. Given the relative infancy of research on inferred motives, this chapter will start with the assumption that these processes are also at work when people make inferences about motives. In addition, we will suggest that each of these processes—simulation and the application of implicit theory—may take one of two forms. With regard to simulation theory, we will propose that simulation can lead to either a passive projection process or a more effortful perspective-taking process. With regard to implicit theories, we will propose that perceivers may rely on either the logic of covariation (Kelley, 1973) or a constraint-satisfaction process (Read & Miller, 1993; Reeder et al., 2004). Although we hope to draw some sharp contrasts between simulation and implicit theories, we also assume that these processes are not mutually exclusive and any given inference about motive may be guided by more than one process.

Using Simulation to Infer Motives

Attempts to understand others often start close to home. Perceivers ask themselves, for example, "What would I be thinking if I were her?" Proponents of the simulation theory suggest that the perceiver "pretends to be in certain states the target is in"(Goldman, 2001), imagining the perceptual input the other person experiences (what the other sees, hears, and touches). The perceiver then tries to experience (or match) the same thoughts and emotions that exist in the target. The results of this simulation are then attributed to the target. In the next section we review some of the evidence for simulation as a general account of how people infer the mental states of others.

Evidence for the Simulation Account

Several lines of evidence are consistent with the idea that perceivers rely on simulation when inferring mental states. First, people commonly report the phenomenological experience of mentally trading places with others in order to imagine their feelings (Van Boven & Loewenstein, 2003). For instance, when watching a contestant on a reality TV show, we might find ourselves wondering "How would I feel if I were the one eating that potato bug?"or "What would motivate me to ever do such a thing?" Second, people apparently project their own transient drive states and social motivations onto others. Van Boven and Loewenstein (2003) told their research participants about a group of hikers who lost their way and had little access to water. Participants who had recently engaged in vigorous exercise (and were thirsty, as a result) were particularly prone to estimate that the hikers would be thirsty (as opposed to hungry). In other words, thirsty people projected their own drive state onto others. Kelley and Stahelsi (1970) reported a similar finding in a study involving the perception of social motivation. Participants who expected to play a prisoner's dilemma game predicted that their partners in the game would have motivations like their own. That is, participants with a competitive orientation tended to expect a competitive partner, whereas participants with a cooperative orientation were more likely to expect a cooperative partner.

A third reason to consider the simulation account is that certain aspects of the process may operate spontaneously. The discovery of *mirror neurons* in monkeys suggests a possible neural basis for the process of entering into another's mental state. When people observe the goal-directed actions of others, such as grasping an object, it is possible that related neurons and hand muscles are activated within the observer (Decety, Chapter 9, this volume; Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995; Gallese & Goldman, 1998). Other research suggests that feelings of empathy may be spontaneously activated in an observer. When people are exposed to someone who is in an emotionally arousing situation, they often spontaneously experience feelings of empathy (Hodges & Wegner, 1997). In turn, empathic feelings, or a feeling of general similarity with the other, may prompt attempts at simulation. For example, a study by Ames (2004) illustrates how feelings of similarity can prompt people to project their own motives onto others. In the course of the research, participants indicated how similar they felt to fraternity members. In another part of the study, Ames asked his participants to read a story about a fraternity member who met a young woman at a dance and then asked her to leave with him. Given the ambiguous nature of the story, participants could construe the fraternity man as either motivated to have casual sex with the woman or as motivated to learn more about the woman's personality and opinions. The results of the study indicated that participants who felt similar (as opposed to dissimilar) to fraternity members were more likely to project their own motives onto the character in the story (e.g., those who had an interest in casual sex tended to perceive that same motive in the character). In sum, the studies reviewed above suggest that people may engage in a simulation-like process when judging the motives of another, particularly when they perceive some similarity between themselves and that person.

Two Forms of Simulation: Simple Projection versus Effortful Perspective Taking

Up to this point, we have used the term "simulation" to refer to the general process by which perceivers mentally trade places with another. In this section we will suggest that this general process may take one of two forms, depending on the extent of cognitive effort expended by the perceiver. When perceivers make their judgments quickly, without expending much effort, the result tends to be a simple form of projection.

In simple projection, the perceiver makes little effort to imagine the situation as experienced by the target person. Instead, the perceiver relies on his or her own perspective and merely projects the feelings that arise from it onto the target. For instance, when Marie Antoinette was informed that the peasants in France had no bread, she recommended that they eat cake. Likewise, thirsty athletes who project their own drive states onto others appear to be engaging in simple projection (Van Boven & Loewenstein, 2003). Such projection seems to represent a failure of what we ordinarily mean by "perspective taking." So it seems important to distinguish such simple projection from the more effortful form of perspective taking described below.

We propose that effortful perspective taking involves an active process in which the perceiver imagines the similarities and differences between his or her own experiences and those of the target person. The perceiver tries to conceptualize the target's situation as it would appear to the target, seeking to appreciate the target's current state of mind. This process may require that the perceiver make some adjustment or accommodation to his or her own mental state before it is projected onto the target. For example, perceivers in the Van Boven and Loewenstein (2003) studies might think, "Well, I am not thirsty now, but I sure would be if I were lost and hadn't had anything to drink all day." Such effortful perspective taking may require a degree of effort and motivation on the part of the perceiver. Next we review evidence for the distinction between simple projection and effortful perspective taking.

We begin by noting that some forms of perspective taking develop at an early age. Children as young as 4 and 5 years old can distinguish between their own belief and that of another person (who has a false belief). For example, if a person believes a physical object is in location *A*, whereas the child knows the object is in location *B*, the child will predict that the person will look for the object in location *A* (Wellman, Cross, & Watson, 2001). In other words, even young children are capable of viewing the world from the perspective of another. Given that adults possess at least a similar capability, we might expect them to routinely demonstrate perspective taking. As described below, however, there is a difference between having the capability to engage in effortful perspective taking and actually doing so in everyday life.

For example, Keysar and his colleagues have identified some disturbing limitations in the perspective taking of adults (Barr & Keysar, Chapter 17, this volume; Keysar, Barr, & Horton, 1998). Starting from a developmental framework, Keysar and colleagues (1998) initially expected to find that adult perceivers used language in a non-egocentric way, putting themselves in the "other person's mind" when they interpret the meaning of language. To their surprise, their data forced them to conclude that adults often behaved egocentrically. For example, in one study perceivers read a story in which Jane recommended an Italian restaurant to David. He tried the restaurant and hated it, leaving a note to Jane that simply said, " . . . it was marvelous, just marvelous." When asked how Jane would understand the note, perceivers in this condition routinely thought that Jane would appreciate the sarcasm in the note (i.e., that David hated the restaurant). Perceivers in this study apparently suffered from a "curse of knowledge," such that they assumed that Jane would have the same privileged knowledge that they possessed. In sum, although perceivers may be quick to project their own understanding of events onto others, they may be slower to appreciate the limitations of another person's perspective. As described below, effortful, non-egocentric perspective taking may occur only under ideal circumstances, requiring motivation and cognitive resources from the perceiver.

Keysar and his colleagues (1998; Epley, Keysar, Van Boven, & Gilovich, 2004) suggest that there may be separate time courses for simple projection as opposed to effortful perspective taking. They propose a monitoring-and-adjustment model as a way of understanding the data on language processing. Accordingly, perceivers initially respond to information in a simplistic, egocentric manner (without regard to the perspective of others). But given sufficient time, perceivers may correct their error and begin to describe the situation in terms of the other person's point of view. In describing how perceivers project their drive states onto others, Van Boven and Loewenstein (2003) also describe a two-stage model. The first stage involves people's predictions of how they would feel in the target person's situation, whereas the second stage involves an adjustment whereby people accommodate perceived differences between themselves and others. This second stage appears more effortful and would appear to require greater motivation on the part of the perceiver.

The research discussed above may have important implications for the simulation account of motive inferences. The evidence suggests that an effortful version of simulation may not be the default process for inferring motives but instead takes place only under ideal circumstances. Indeed, Kawada and colleagues (2004) found that people tend to project their goals onto others automatically, without having to consciously place themselves in another person's shoes. If perceivers subsequently engage in a more conscious and controlled effort to understand others' motives, the initial projections of motive would presumably undergo some alteration. Future research might aim to delineate the circumstances when people are most prone to engage in simple projection versus effortful perspective taking. For example, it may be the case that when the perceiver and target person are presented with the same stimulus, or are in the same situation, perceivers are most prone to simple projection. Although we recognize that distinguishing between simple projection and effortful perspective taking is not without difficulty, we believe that future researchers will find the effort worthwhile.

We conclude this discussion by noting that our aim here is to draw a distinction at the level of process rather than in terms of outcome or accuracy. For example, a father may think that his son wants to go to college simply because the father would like to have had the opportunity to go to college ("I wish I could have 4 years to learn about the world and explore my talents. I am sure Junior feels the same!"). The example is a case of simple motive projection. On the other hand, if the father actively places himself in his son's shoes and tries to imagine the world as it would appear from that unique vantage point, the process is one of effortful perspective taking ("Let's see . . . Junior likes tail-gate parties and cheerleaders, and there are lots of both in college"). Although effortful perspective taking is certainly a more sophisticated and potentially accurate way to infer motives, it may not always lead to accuracy (e.g., Junior might not want to go to college because he hates to study, or, alternatively, Dad might be mistaken in his assumptions about Junior's likes and dislikes). In the next section we will consider an alternative approach to inferring motives.

Using Implicit Theories to Infer Motives

In addition to using simulation, perceivers may employ abstract, theory-like constructs to understand the mental states of others (Gopnik & Wellman, 1994). Such theories are implicit in the sense that people may not be able to articulate the details of the theory, but the existence of a coherent theory can be inferred indirectly from the kinds of judgments that people make. We will attempt to identify some patterns in people's judgments about motives, and we offer some suggestions about the nature of the underlying theory and process. Although there is fair agreement among developmental psychologists and philosophers that a theory of mind exists, these literatures are less specific about the concrete assumptions and psychological processes of implicit theories. We will draw from the social psychological literature to specify two general types of theoretical reasoning about motives. The first of these involves the principle of covariation popularized by attribution theorists (Jones & Davis, 1965; Kelley, 1973), whereas the second involves constraint-satisfaction processes that are implicit in the early person perception literature (Heider, 1958) and explicit in more recent theorizing (Read & Miller, 1993; Reeder et al., 2004).

Covariation and Inferences about Motives

Heider (1958, p. 152) noted a fundamental pattern in attribution: a potential cause (or motive) "will be held responsible for an effect which is present when the effect is present and which is absent when the effect is absent." Heider noted that this covariation principle underlies Mill's method of experimental inquiry. The principle suggests that if black clouds are present when it rains and absent on days when it is dry, people will tend to see black clouds as the cause of rain. Kelley (1973) elaborated on the covariation principle in order to explain the circumstances under which perceivers will make global causal attributions to internal (vs. external) causes.

The principle also underlies the landmark "noncommon effects" analysis offered by Jones and Davis (1965). Although the noncommon effects analysis is often interpreted as a theory of trait attribution, it actually describes how motives are inferred. Jones and Davis illustrated their theory with a quaint example of Ms. Adams, who received multiple marriage proposals from eligible suitors. The analysis begins by noting that any choice on her part typically is consistent with multiple motives and that the perceiver will try to select among these possible motives. Suppose that two of her suitors—Mr. Bagby and Mr. Caldwell—are both handsome and wealthy. Yet, Bagby holds a prominent social position, whereas Caldwell does not. If Ms. Adams selects Bagby as her marriage partner, covariation logic suggests that we can rule out good looks and wealth as the primary motives for her choice because they are held in common by both suitors (Sutton & McClure, 2001). Yet, her selection does covary in a unique way with Bagby's social position. Consequently, we are likely to infer that Adams's choice was determined by that unique motive. In other words, motives that are unique to her choice are considered to be the more plausible ones.

But what if there are two or more unique motives underlying a particular action? Suppose that the lucky man, Bagby, not only holds a prominent social position but also possesses a great sense of humor. Kelley's discounting principle suggests that perceivers will discount (or downplay) any particular motive if there are other plausible motives that might account for the behavior. According to the discounting principle, then, perceivers should be more certain about Ms. Adams's motivation if Bagby is merely socially connected, as opposed to being both socially connected and a humorous fellow. In short, the principles of covariation and discounting suggest that perceivers will search for a particular motive that covaries in some unique way with a person's choices and actions.

Constraint Satisfaction and Inferences about Motives

In general, constraint satisfaction deals with finding a "fit" among various elements in a system, as when a university administrator develops a class schedule by arranging classrooms, meeting times, professors, and students into the schedule (Thagard, 1996). Earlier, we noted that the logic of covariation involves isolating a motive that is unique to a particular action. In contrast, the logic of constraint satisfaction involves finding a motive that fits with, or is common to, a variety of actions. Heider (1958, p. 51) suggests that there is an economy to perceivers' logic such that they look for a motive that helps to reconcile apparent contradictions in a person's behavior. For example, suppose perceivers learn about a young man who lavished attention on an elderly woman, received expensive gifts as a result, and then terminated the relationship. Perceivers are likely to attribute a mercenary motive to the young man (the term "gold digger" comes to mind). The mercenary motive accounts for the full pattern of behavior.

If a variety of motives are implied by a target person's behavior, perceivers sometimes select more than one motive. In this case, perceivers seek consistency among the selected motives (Heider, 1958, p. 52). For instance, in the no-choice condition of the study described earlier by Reeder and colleagues (2004), participants learned that a student worker had received instructions from her superior to help professors move books. After watching the student help the professor, participants inferred that the student was motivated by both obedience and helpfulness. Apparently, the motive to obey was perceived as consistent with a helpful motive. Evaluative consistency is, perhaps, the major basis for such consistency. Being obedient on the job and being helpful both have a positive valence. Consequently, perceivers may not hesitate to endorse both motives simultaneously. In contrast, if two potential motives are evaluatively inconsistent, perceivers are likely to reject one of them in favor of the other. In a second condition of the Reeder and colleagues study, the student's helpful behavior was portrayed as possibly due to an ulterior motive (wanting to win an award for which the professor was responsible). In this condition, a helpful motive would appear inconsistent with the negative valence of the ulterior motive. Consequently, a constraint-satisfaction process would predict that perceivers would attribute lower ratings of helpfulness to the student in this condition. This prediction was confirmed. Thus, it appears that perceivers are unlikely to attribute motives with contrasting valences. When perceivers *do* attribute more than one motive to a person, the motives tend to fit together in a consistent manner.

It is worth noting that the constraint-satisfaction process sometimes leads to predictions that are at odds with a simple covariation analysis. Unlike covariation, constraint satisfaction implies that the presence of two potential motives for an act does not automatically cast doubt on one of them. For instance, in the example cited earlier where a student worker helped a professor out of apparent obedience, the presence of the obedience motive did not lower participants' ratings of helpfulness. Whether or not a given motive will be discounted, then, depends less on the number of other plausible motives than on the "fit" between the possible motives.

Before concluding our discussion of constraint satisfaction, we think it is important to note that the process can introduce considerable bias. When perceivers have prior knowledge of a target person, either because of earlier encounters or stereotypical expectations (Ames, 2004), they are likely to infer a motive that is consistent with their expectations. Thus, perceivers who dislike a politician are likely to see dark motives in even the most benign actions of the politician. This tendency may be particularly strong because people's motives are typically ambiguous, allowing for any number of interpretations.

In general, then, perceivers may be biased by their preconceptions, or by their own unique goals and perspective, to spot particular types of motives in another person. Heider (1958, p. 172) suggested that perceivers choose a reason for a person's actions that (1) fits the perceiver's wishes and (2) fits the data. Below we will describe naive realism (Ross & Ward, 1996) and self-esteem enhancing tendencies (Tajfel & Turner, 1986; Taylor & Brown, 1988) as possible biasing factors in the process of attributing motives.

Naive realism is said to represent the last vestige of the kind of egocentrism found in children (Ross & Ward, 1996). This sort of thinking consists in believing that one perceives events objectively (as they really are) and that other rationally minded individuals will see them similarly. If others fail to see things similarly it is because they are lazy, irrational, or biased by ideology or self-interest. This last tendency—to see others as biased by self-interest—is particularly relevant to our discussion of perceived motives. Naive realism suggests that when another person (or group) disagrees with our opinions we tend to see the motive of self-interest at work ("They have an axe to grind!"). Indeed, Reeder, Pryor, Wohl, and Griswell (in press) found such a bias in a survey of Americans and Canadians regarding their support for the 2003 war in

Iraq. People on both sides of this controversial issue tended to see those on the other side as guided by self-interest (as opposed to being guided by ethical principles). Thus, although people believe the motive of self-interest is widespread (Miller, 1999), they are particularly prone to see it in others who disagree with their opinions.

Finally, identity concerns and self-esteem needs (Tajfel & Turner, 1986; Taylor & Brown, 1988) may also lead people to make biased attributions of motive (Reeder et al., in press). When other individuals or groups agree with our own position on a controversial issue, it not only validates our view of the world, it validates us. Consequently, we feel more positive about those who support our opinions, and we tend to attribute positive motives to them. Similarly, self-esteem needs may lead us to attribute positive motives to a subordinate who flatters us, whereas an observer (who overhears the flattery) might attribute an ulterior motive to the ingratiator (Vonk, 2002). In sum, both naive realism and the tendency to self-enhance may lead to biased attributions of motive.

CONCLUSION

The evidence reviewed in this chapter suggests that inferences about motives are a common occurrence in everyday life. Perceivers often infer motives spontaneously and rely on them to better understand the people in their lives. Inferences about motives appear to be an especially important determinant of trait attributions. Such findings challenge traditional attribution theories that view abstract causal attributions as underlying the trait attributions we make about others. Although little is currently known about how people infer motives, we suggested that processes related to simulation and implicit theory may be important. With regard to inferring motives via simulation, we drew a distinction between simple projection (i.e., projecting one's own motives onto others without considering their unique circumstances) and a more resource-dependent form of effortful perspective taking. With regard to using implicit theories to infer motives, we suggested that such theories might operate via principles of covariation analysis and constraint satisfaction.

From the summary above, it is apparent that we covered a lot of ground in this chapter. Nevertheless, our topic is still largely unexplored, and many questions remain. In particular, we hope future research will shed light on the similarities and differences between inferences about motives and inferences about other mental states such as beliefs and feelings.

REFERENCES

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299–352.

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

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187–276.

Dennett, D. C. (1993). Three kinds of intentional psychology. In S. M. Christensen & D. R. Turner (Eds.), *Folk psychology and the philosophy of mind* (pp. 121–143). Hillsdale, NJ: Erlbaum.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*, 327–339.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology, 73*, 2608–2611.

Fiedler, K., & Schenck, W. (2001). Spontaneous inferences from pictorially presented behaviors. *Personality and Social Psychology Bulletin, 27*, 1533–1546.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*, 493–501.

Goldman, A. I. (2001). Desire, intention, and the simulation theory. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 207–224). Cambridge, MA: MIT Press.

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). Cambridge, UK: Cambridge University 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.

Hodges, S., & Wegner, D. M. (1997). Automatic and controlled empathy. In W. J. Ickes (Ed.), *Empathic accuracy* (pp. 311–339). New York: Guilford Press.

Idson, L. C., & Mischel, W. (2001). The personality of familiar and significant people: The lay perceiver as a social-cognitive theorist. *Journal of Personality and Social Psychology, 80*, 585–596.

Johnson, J. T. (2001). On weakening the strongest link: Attributions and intervention strategies for behavior change. *Personality and Social Psychology Bulletin, 27*, 408–422.

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.

Kawada, C. L. K., Oettingen, G., Gollwitzer, P. M., & Bargh, J. A. (2004). The projection of implicit and explicit goals. *Journal of Personality and Social Psychology, 86*, 545–559.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28*, 107–128.

Kelley, H. H., & Stahelski, A. J. (1970). The inference of intention from moves in the prisoner' dilemma game. *Journal of Experimental Social Psychology, 6*, 401–419.

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science, 7*, 46–51.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3*, 23–48.

Malle, B. F., Knobe, J., O'Laughlin, J. O., Pearce, G. E., & Nelson, S. E. (2000) Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *Journal of Personality and Social Psychology, 79*, 309–326.

McClure, J. (2002). Goal-based explanations of actions and outcomes. In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology, 12*, 201–235.

Miller, D. T. (1999). The norm of self-interest. *American Psychologist, 54*, 1053–1060.

Nussbaum, S., Trope, Y., & Liberman, N. (2003) Creeping dispositionalism: The temporal dynamics of behavior prediction. *Journal of Personality and Social Psychology, 84*, 485–497.

O'Laughlin, M. J., & Malle, B. F. (2002). How people explain actions performed by groups and individuals. *Journal of Personality and Social Psychology, 82*, 33–48.

Read, S. J., & Miller, L. C. (1993). Rapist or "regular guy": Explanatory coherence in the construction of mental models about others. *Personality and Social Psychology Bulletin, 19*, 526–541.

Reeder, G. D., Hesson-McInnis, M., Krohse, J. O., & Scialabba, E. A. (2001). Inferences about effort and ability. *Personality and Social Psychology Bulletin, 27*, 1225–1235.

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology, 83,* 789–803.

Reeder, G. D., Pryor, J. B., Wohl, M. J. A., & Griswell, M. L. (in press). On attributing negative motives to others who disagree with our opinions. *Personality and Social Psychology Bulletin.*

Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J. & Lawrence, M. (2004). Dispositional attribution: Multiple inferences about motive-related traits. *Journal of Personality and Social Psychology, 86*, 530–544.

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. E., Reed, E. Turiel, & T. Brown (Eds.), *Values and knowledge* (pp. 103–135). Mahwah, NJ: Erlbaum.

Smith, E. R., & Miller, F. D. (1983). Mediation among attributional inferences and comprehension processes: Initial findings and a general method. *Journal of Personality and Social Psychology, 44*, 492–505.

Shaver, K. G. (1975). *An introduction to attribution processes*. Cambridge, MA: Winthrop.

Sutton, R. M., & McClure, J. (2001). Covariational influences on goal-based explanation: An integrative model. *Journal of Personality and Social Psychology, 80*, 222–236.

Taylor, S. E., & Brown, J. D. (1988*)*. Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193–210.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago: Nelson-Hall.

Thagard, P. (1996). *Mind: Introduction to cognitive science*. Cambridge, MA: MIT Press.

Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin, 29*, 1159–1168.

Vonk, R. (1998). The slime effect: Suspicion and dislike of likeable behavior toward superiors. *Journal of Personality and Social Psychology, 74*, 849–864.

Vonk, R. (2002). Self-serving interpretations of flattery: Why ingratiation works. *Journal of Personality and Social Psychology, 82*, 515–526.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development. The truth about false belief. *Child Development, 72*, 655–684.

White, P. A. (1995). Use of prior beliefs in the assignment of causal roles: Causal powers versus regularity-based accounts. *Memory and Cognition, 23*, 243–254.

8

Explanatory Coherence and Goal-Based Knowledge Structures in Making Dispositional Inferences

STEPHEN J. READ
LYNN C. MILLER

"The tailor carries the old woman's groceries across the street" (in Winter & Uleman, 1984). If you are like most people, upon reading this sentence you made the inference that the tailor was "helpful." Social psychologists have long argued that such trait inferences are central to social perception and social interaction. Moreover, considerable research demonstrates that people make such trait inferences spontaneously (quickly and with no conscious intentions) in response to a wide array of social events (for a review, see Uleman, Newman, & Moskowitz, 1996; see also Carlston & Skowronski, 1994).

But what do such trait inferences have to do with the problem of other minds? A great deal. We, along with several other authors (e.g., Reeder & Trafimow, Chapter 7, this volume; Ames, Chapter 10, this volume) argue that inferences about the goals and mental states of people are central to making trait inferences about them (see also Jones & Davis, 1965). Ultimately, we are making two related claims here: (1) inferences about an actor's goals typically play a central role in making trait ascriptions about them, and (2) trait ascriptions about an actor typically involve making claims about the goals and mental states of the actor.

However, despite the evidence that trait inferences are frequently and spontaneously made and that goal inferences play an important role in trait ascriptions, there are no process models of which we are aware that provide an account of *how* such spontaneous inferences are made. In this paper, we will outline such a process account of trait inferences, the social dynamics model. This account is based on a general model of person perception that we have outlined in greater detail in other places (Read, 1987; Read & Miller, 1993, 1998). Central to this model is the idea that inferences about an actor's goals frequently drive trait inferences.

The social dynamics model argues that we comprehend other people's minds by creating a coherent narrative or story of their actions, organized around their goals (Miller & Read, 1991; Read, 1987; Read & Miller, 1995). This narrative or story represents: (1) the actor's goals and reasons for his or her actions, (2) the conditions that might instigate those reasons, (3) the actor's plans and how they relate to the actor's goals and reasons, and (4) the consequences or outcomes of the actor's actions.

Our model is based on two key assumptions. The first assumption is that understanding others relies on the use of detailed, goal-based social knowledge structures (Read, 1987; Schank & Abelson, 1977) that are activated during the process of narrative construction. As we will argue in greater detail, traits are one such goal-based structure. The second key assumption is that central to the construction of these narratives are processes of explanatory coherence that derive from constraint satisfaction processes in neural networks (Miller & Read, 1991; Read & Miller, 1993; Thagard, 1989, 2000; see also Kintsch, 1998). These constraint satisfaction processes act to organize the activated knowledge structures into a coherent narrative that provides the best explanation of the other's actions.

TRAIT INFERENCES

Our model is intended as a general model of person perception and explanation (Read & Miller, 1998); accordingly, it provides a detailed processing account of how trait inferences are made in response to social events, such as the one that opened this paper. A key point of the model is our assumption about how trait concepts are cognitively represented. We argue that traits are frame-based structures or schemas in which the goals of the actor play a central representational role. One important point that follows from the assumption that goals are central to the representation of traits is that inferences about the goals of an actor should then play a central role in dispositional inference (Read, Jones, & Miller, 1990; Read & Miller, 1989).

Over the past 15 years, several different researchers (e.g., Borkenau, 1990, 1991; Fleeson, Zirkel, & Smith, 1995; Read, Brownstein, & Miller, 2005; Read, Jones, & Miller, 1990; Read & Lalwani, 1998) have provided evidence that traits are goal-based categories (Barsalou, 1992) in which the goals associated with a trait are central to the trait's meaning. For example, Read, Jones, and Miller (1990) showed that each of six major traits (e.g., Dominant, Extroverted, Agreeable) in Wiggins's (1979) Interpersonal Circumplex was strongly associated with a different, small set of goals. Further, they showed that ratings of the extent to which each of a set of behaviors achieved the goals associated with each trait strongly predicted ratings of how typical the behavior was of someone with that trait (see also Borkenau, 1990, 1991; Fleeson, Zirkel, & Smith, 1995).

More recently, Read and colleagues (2005) conducted a large-scale web survey in which participants were asked to indicate which goals were descriptive of people characterized by each of 55 traits (11 were sampled from each of the Big Five trait dimensions). Participants strongly agreed in identifying a small set of central goals for each trait.

Our assumptions about traits go beyond simply assuming that traits are lists of features, such as the goals associated with the trait. Rather, we argue that traits are structured, organized schemas or frames that encode the sequence of components that count as the behavioral exemplars of each trait (Miller & Read, 1991; Read & Miller, 1998; see also John, 1986). One can think of traits as little "stories" or "narratives" that encode the goals of the actor, the factors that instigated the actor's goals, the actions the actor takes in response to those goals, and the consequences of the actions. For example, consider the trait *helpful*. When people ascribe the trait *helpful* to Person B on the basis of his or her actions, the following is typically true of the social interaction:

Person A has a need, which leads to
Person B noticing that need, which instigates
Person B forming an intention to help Person A, which provides the reason for
Person B taking an action intended to help Person A, which results in
Person A being helped by Person B's action.

Even though one may point to instances in which an ascription of *helpful* is made in the absence of one or more of these components, the frame above is the prototypical or "ideal" structure for the trait *helpful*. Each of these components is important to the characterization of someone as helpful. This analysis of the *helpful* frame will play a central role in our subsequent account of how people make trait inferences of *helpful* from such events or descriptions as the one with which we began this chapter.

Another interesting trait frame is that for *vindictive*. When people say that Person A is Vindictive they are typically claiming that the following is true:

> Person B did something to hurt Person A, which leads to
> Person A thinking the hurt was intentional, which instigates
> Person A's motivation for revenge, which provides the reason for
> Person A's hurtful actions toward Person B, which results in
> Person A hurting Person B.

One of the things that makes *vindictive* particularly interesting is that it actually is constructed of *two* little stories, one of which leads to the other:

> Person B intentionally hurts Person A, which leads to
> Person A having the goal of hurting Person B and therefore taking steps to hurt Person B.

A similar analysis can be provided for a number of different trait terms.

Consistent with the centrality of goals in the representation of traits, several researchers have demonstrated that inferences about an actor's goals play an important role in trait inferences about an actor. Consider, for example, a study by Read, Jones, and Miller (1990). They took the six traits and the 100 behaviors generated for each trait from Buss and Craik's (1983) act frequency theory of personality. They then (1) identified two goals central to the representation of each of the six traits, (2) had subjects judge how related each of the 100 behaviors was to the goals associated with that trait, and (3) for each behavior had participants rate how strongly they would infer the target trait from each behavior. They found that the goal-relatedness of a behavior strongly predicted the strength of the trait inference from the same behavior.

Read and Lalwani (1998) used a different approach to provide further evidence for the role of goals in trait inferences. They created a series of two-sentence stories in which each pair of sentences could be put in two different orders that resulted in two very different narratives. The two narratives led to two very different sets of inferences about the goals of the central actor in the story and thereby led to different trait inferences. Consider this example from the study:

> *Order 1*
>
> Barry told his pals that he had an original idea for improving the economy and the environment—to create jobs in recycling. The next day Barry was watching TV with his dad, and he heard President Clinton say that recycling jobs could potentially help both the environment and the economy.
>
> *Order 2*
>
> Barry was watching TV with his dad, and he heard President Clinton say that recycling jobs could potentially help both the environment and the economy. The next day Barry told his pals that he had an original idea for improving the economy and the environment—to create jobs in recycling.

These two orders result in very different inferences about Barry's goals and his traits. To experimentally test this hypothesis, for each set of sentences we developed a set of three goals and three traits that we thought would be consistent with one story and three goals and three traits that we thought would be consistent with the alternative story. For example, for order 1, one goal was to be politically involved, and one trait was *socially conscious*. In contrast, order 2 should lead to different goal and trait inferences (e.g., goal: to gain popularity; e.g., trait: *phony*).

Different groups of subjects were given the two orders of sentences for each story. After reading each story they rated both the goals of the actors and their traits. As predicted, subjects reading the different orders made very different inferences about the actors' goals, as well as making very different trait inferences about the actors. Further, the participants' inferences about the actors' goals strongly predicted their inferences about the actors' traits. And once we statistically controlled for the goal inferences, the impact of the different story orders on trait inferences was almost eliminated. This supports the hypothesis that goal inferences play an important role in trait ascriptions about an actor.

Daniel Ames has also provided evidence for the importance of goals in impressions. For example, Ames, Flynn, and Weber (2004) showed that inferences about whether people were helpful were partially mediated by inferences about their underlying goals. Less directly related to the current point, they have also shown that goal inferences mediate the link between overt behaviors and general impressions (good/bad).

Finally, Glenn Reeder and his colleagues (Reeder, Hesson-McInnis, Krohse, & Scialabba, 2001; Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002; Reeder, Vonk, Ronk, Ham, & Lawrence, 2004) have reported a number of studies showing that motive inferences are central to trait inferences. For example, in various studies, they have shown that situational constraints can promote trait inferences—in contradiction to Jones and Davis's (1965) claim that situational constraints inhibit trait inferences—when the situation elicits a particular motivational or mental state inference that forms the basis for a trait inference. For example, someone who blows a soccer kick because he was paid to do so is considered to be dishonest.

SOCIAL DYNAMICS MODEL

In the preceding section, we have outlined a frame-based account of trait representations. Now we describe a process model of how these trait frames are used in the dispositional inference process. Read and Miller's (1998) social dynamics model describes a feedback neural network that provides an account of how human actions are assembled into coherent stories or narratives and how dispositional inferences are made from those narratives.

Figure 8.1 provides an overview of the conceptual organization of the model. There are four hierarchically arranged layers with feedback links among layers and within layers. In conceptualizing how perceivers take information from the environment and ultimately make a trait inference about an actor, it is useful to break the necessary information processing into these four layers. (One could break the processing down further, but four layers are sufficient to capture what we view as the central distinctions.)

Concepts lower in the hierarchy send activation to higher-level concepts, and conversely higher-order concepts can influence the activation of lower-level concepts. So, for example, at the lowest level various features of an individual, such as carrying a clothing tape measure and coming out of a tailor shop might activate the higher-level concept of tailor. On the other hand, once people have used these features to identify someone as a tailor they might be more likely to "see" him holding a needle and thread. At the same time, concepts within a layer can also influence one another: For example, seeing that the tailor is carrying a needle might make it more likely that people will "see" him carrying a thread, as well.

One important feature of this model is that activated concepts can activate related information. For example, the concept "old woman" probably activates related stereotypic information, such as frail, needing help, and so on.

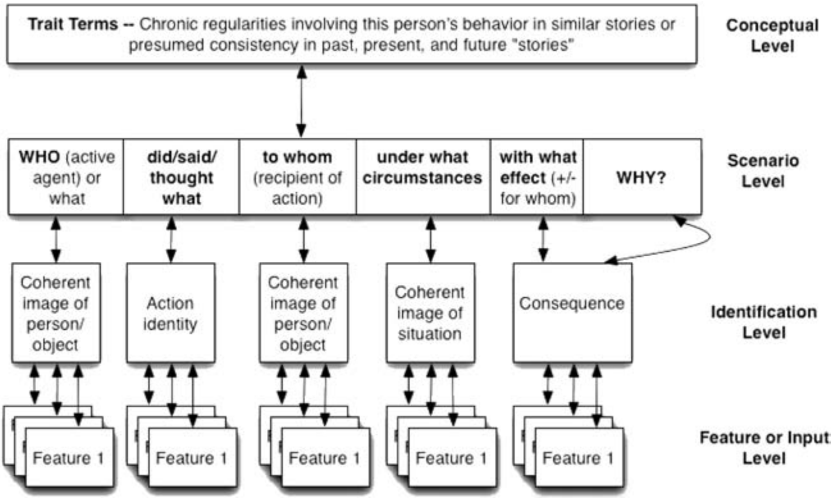
Trait Terms -- Chronic regularities involving this person's behavior in similar stories or presumed consistency in past, present, and future "stories"
Conceptual Level
WHO (active agent) or what
did/said/ thought what
to whom (recipient of action)
under what circumstances
with what effect (+/- for whom)
WHY?
Scenario Level
Coherent image of person/ object
Action identity
Coherent image of person/ object
Coherent image of situation
Consequence
Identification Level
Feature 1
Feature 1
Feature 1
Feature 1
Feature 1
Feature or Input Level

FIGURE 8.1. The conceptual structure of the social dynamics model of person perception and dispositional inference.

The input or feature layer receives input from the environment about the features of actors, objects, and actions. Activation from the features then goes to the identification layer, where the features are used to identify particular actors, such as a tailor or an old woman, objects such as groceries, street signs, curbs, etc., and actions, such as carrying or walking.

Once the various actors, objects, and actions are identified, they must then be assembled into a representation of the social interaction or scenario. This happens in the scenario layer, where various actors, objects, and actions are assembled into a coherent scenario that identifies the goals and motives of the actor. The results of the scenario layer then proceed to the meaning or conceptual layer, where trait identifications are made from the scenarios. As we move higher in the hierarchy, processing is more dependent on available cognitive capacity.

To better understand how the model works, it is useful to work through an example of a trait inference. Let us take the sentence with which we began the chapter and work through what our model suggests is the process by which such a spontaneous trait inference is made. In working through the example, let's imagine for a minute that, rather than reading the sentence, people actually observe the interaction described by the sentence "The tailor carries the old woman's groceries across the street."

Identification of Actors, Objects, Actions

The first step is to identify the actors, the objects, and the actions. Features of the actors, objects, and actions that are activated in the input or feature layer by input from the environment will send activation to linked concepts in the identification layer. Thus, various features of the tailor, such as coming out of a tailor shop, carrying pins and needles, and so on, will activate the concept "tailor" in the identification layer. Physical features of the "old woman," such as gender cues and cues to age, such as wrinkled skin, white hair, and slow gait, would allow us to identify her as an "old woman." Physical features of the "groceries," such as the brown paper bag or the celery stalks peeking out of the top, will allow us to identify the groceries. And physical actions by the tailor, such as supporting the bag and walking across the street, identify the concept of "carrying." Because this is a feedback network, concepts can feed activation back to features. Thus, activation of a concept such as "groceries" may make it more likely that people will "see" the tailor carrying milk. Furthermore, activated concepts can send activation to one another, thereby supporting or inhibiting one another. For instance, alternative identifications of an individual, such as whether the man is a tailor or a store clerk, can inhibit one another.

Scenario Level: Creating a Scenario

Once the actors, objects, and actions are identified, they must be assembled into a particular scenario, or "narrative." The action, or the verb that represents the action, is represented by a particular frame or case with relevant roles or slots. These assumptions about the representation of action follow from work on verbs (Barsalou, 1992; Fillmore, 1968; Jackendoff, 1983) and from work on text comprehension (e.g., Graesser & Clark, 1985; Kintsch, 1998). Roles or slots correspond to things such as the performer of an action, the recipient (if any) of an action, and the objects involved in the action. The various roles or slots have associated with them specifications of the attributes of the kinds of things that fit into the slots.

Scenario Level: Why?

Once people have created a scenario from the actors, objects, and actions, they can then try to identify the goals and reasons for the action, if they exist. We suggest that the assembled scenario will activate in memory a number of possible related goals and reasons: for example, wanted to help, wanted to get the old woman to like him, expected to be paid, knew that his wife was watching and expected him to help, and so on.

But once all these alternative reasons are activated, how do people choose among the alternatives? We have argued that constraint satisfaction processes, which arise naturally out of the mutual interactions among elements in a feedback neural network, will act to choose from alternative explanations (Read & Miller, 1993, 1998). In the current context, goals and reasons are one very important class of explanations for human behavior.

Constraint satisfaction processes actualize a set of principles of explanatory coherence (Thagard, 1989, 2000) that are implemented in terms of patterns of connectivity among explanations and things to be explained. Consistent explanations or explanations that support one another have excitatory links, whereas competing or alternative explanations have inhibitory links. In this kind of network, when two explanations compete, the stronger explanation will inhibit the weaker one.

A constraint satisfaction network realizes the following principles of explanatory coherence (Read & Marcus-Newhall, 1993; Thagard, 1989, 2000):

- *Breadth:* Explanations that explain more of the behaviors will receive activation from more concepts and thus will be more strongly activated.
- *Simplicity:* Activation is divided among explanations, such that if more explanations are needed each will receive less activation and be viewed as weaker.
- *Recursive explanation:* Events in a causal chain receive activation from things they explain and things that explain them. Thus, explanations that are part of a causal chain will receive higher levels of activation.

Furthermore, we have noted elsewhere (Read & Miller, 1993) that these constraint satisfaction processes also implement two other central principles of explanation: *discounting* and *augmenting*. We can see how augmenting might play a role in this example. Given that the man was a tailor and that helping the old woman was not part of his role, people might infer that he helped in spite of his role, which might lead to a stronger inference of *helpful*.

Read and Marcus-Newhall (1993) and Read and Lincer-Hill (1998) have provided empirical evidence for the role of these principles of explanatory coherence in both social explanation and dispositional inferences. Thagard (1989, 2000; Kunda & Thagard, 1996) has provided many examples of the application of a computational implementation of constraint satisfaction principles (Explanatory Coherence by Harmany Optimization, or ECHO) to a wide variety of scientific, legal, and everyday reasoning tasks (see also Read & Miller, 1993; Read, Vanman, & Miller, 1997).

Conceptual Level: Trait Inferences

To summarize, people have now assembled a scenario from the actions and have identified the possible goals and reasons of the actors in that sequence. How do they make trait inferences from this? As argued above, traits are represented as frame-based structures that identify the central actions of a sequence of behaviors and the goals of and reasons for that sequence. We suggest that components of the meaning representation will send activation to the corresponding components of trait frames. Thus, traits can be retrieved and used to characterize a behavior by matching features of the action's representation against the feature representation of various traits.

For example, the components of the meaning representation constructed for the action *The tailor carried the old woman's groceries across the street* match each aspect of the trait frame for *helpful*:

- The old woman needed help carrying her groceries, which led to
- The tailor noticing that need, which instigates
- The tailor forming an intention to carry the old woman's groceries, which provides the reason for
- The tailor taking an action intended to help the old woman (carrying her groceries), which results in
- The old woman being helped by the tailor's actions (the groceries are carried across the street).

As a result, the trait frame for *helpful* should be highly activated. Note that inferences about the goals and intention of the actor are central to the activation of the trait frame. If the trait frame is highly activated and succeeds in inhibiting alternatives, then this trait should be used to characterize the actor. Thus, this model provides an account of how trait inferences are made from observed behavior by creating a coherent narrative in which the goals of the actor are central.

Other Applications of This Approach

Both Reeder and Trafimow (Chapter 7, this volume) and Ames (Chapter 10, this volume) suggest that coherence-based processes may apply to some of the findings they discuss. Here we outline in more detail how our coherence-based model might apply to some of their findings.

Ames

One major question that Ames (Chapter 10, this volume) is concerned with is the following: When is person perception influenced by social projection, and when is it influenced by stereotypes? Two principles of our model are relevant to this question. One has to do with which set of concepts is more likely to be activated by input features. In a model such as our social dynamics model, the activation of higher-order concepts will depend on the number of relevant input features that are activated. Greater overlap between the input features and the concepts leads to higher activation of the concepts.

Ames has demonstrated that social projection is more likely when the target is viewed as more similar to the perceiver (i.e., has more features in common). Such similarity should result in greater overlap with self-related concepts and thus lead to greater activation of the self-related concepts. On the other hand, greater similarity (i. e., more shared features) between the target and a stereotype should lead to greater activation of the stereotype. Thus, one factor that should strongly influence the extent to which social projection or stereotype use takes place is the relative degree of similarity between the target and the two relevant sets of concepts (perceiver and stereotype), and their comparative levels of activation.

A second important factor is the nature of the relationship between the self-related concepts and the stereotype concepts: Are they inconsistent with each other, unrelated to each other, or consistent with each other? Inconsistent concepts should inhibit each other, whereas consistent concepts should mutually reinforce each other. One implication of our analysis is that under some circumstances social projection and stereotype use could simultaneously influence mindreading or perhaps even reinforce each other (when the concepts are positively linked), whereas under other circumstances they would compete for influence (when the concepts are negatively linked) and would provide alternative sources of influence.

Reeder

Reeder and Trafimow (Chapter 7, this volume) note that the classic models of dispositional inference (e.g., Gilbert, 1998; Jones & Davis, 1965; Trope, 1986) assume a hydraulic relationship between dispositional and situational causes. However, contrary to this assumption, Reeder and his colleagues (Reeder et al., 2001, 2002; Reeder, Prior, Wohl, & Griswell, 2004) have shown that situational constraints can frequently lead to strong dispositional inferences, evidently because the situational factors induce motivational inferences, which subsequently lead to strong trait inferences.

For example, in one study, a student named Sara helped a professor move a stack of books. In the free-choice condition, there was no evident pressure on her: she freely chose to help move the books. In the no-choice condition, Sara's supervisor instructed her to help professors with tasks. Finally, in the ulterior-motive condition, participants were told that she might want to impress the professor because it would help her win an award. In line with Reeder's multiple inference model, participants in the free-choice condition rated her as a helpful person. More important, in the no-choice condition, participants rated her as obedient, because she followed her supervisor's instructions, and they also rated her as a helpful person. Reeder argues that this outcome occurred because obedient and helpful are positively linked; people who are obedient are also viewed as likely to be helpful. (Note, however, that this pattern of trait inference is strongly inconsistent with the classic, hydraulic model of person versus situation.) Finally, in the ulterior-motive condition, participants rated Sara as selfish, but not at all helpful. Reeder argues that one of the reasons she is rated so low on helpfulness in this condition is because selfishness and helpfulness are inconsistent with each other.

The reader should already see where our analysis is going. We can use the general framework we have outlined here, including our conceptual analysis of the trait *helpful*, to explain how situational cues can actually lead to greater dispositional inference in this situation. The analysis in the free-choice condition is fairly straightforward. Sara's behaviors in aiding the professor to stack books fit the frame for the trait *helpful*, and the available information does not activate any competing interpretation. Thus, she is likely to be viewed as a helpful person.

The other two conditions are somewhat more complicated. In the no-choice condition, Sara's behavior fits the frame for *helpful*, but because the behaviors are in response to instructions from her supervisor, the behaviors also fit the frame for *obedient* (the short form of the frame would be doing what you are told to do). Thus, initially two possible interpretations of the behavior and two possible trait inferences are activated. Our social dynamics model suggests (as does Reeder) that what then happens depends on the relation between the two activated traits. If they are consistent with each other (have an excitatory relationship), as seems reasonable and as Reeder's data suggest, then they may mutually support each other. In that case, the perceiver may decide that both traits accurately characterize Sara, which is what was found.

But now consider the ulterior-motive condition. Again, Sara's behaviors would activate the trait frame *helpful*. In addition, the ulterior-motive information, in conjunction with the behaviors, might simultaneously activate something like the trait selfish. However, the traits helpful and selfish are inconsistent with each other; they should have a negative or inhibitory relationship and should inhibit each other. In this particular case it seems plausible that the ulterior-motive information would lead to greater activation of the trait selfish than of the trait helpful and that, as a result, selfish would then successfully inhibit the activation of helpful. The result would be a strong trait inference of selfish and little or no inference of helpful. This is what Reeder and his colleagues found.

CONCLUSION

Here we have used the social dynamics model to provide an account of how knowledge of other minds (e.g., their goals and intentions) plays a central role in dispositional inference. One further implication of this account is that, by placing goal inferences at the center of dispositional inferences, it suggests that individuals with deficits in theory of mind (e.g., autistic individuals, young children) will have considerable difficulty making sophisticated trait inferences about others.

Although our ultimate focus was on dispositional inference, the model also has much to say about how people make inferences about the goals and motives of others. In fact, it is intended as a general model of social perception and impression formation that provides an account of many different facts of social perception. In other places (e.g., Miller & Read, 1991; Read, 2004; Read & Miller, 1993, 1998), we have outlined in much greater detail how this (or a closely related model) can capture a number of different aspects of how people "read" other minds.

REFERENCES

Ames, D. R., Flynn, F. J., & Weber, E. U. (2004). It's the thought that counts: On perceiving how helpers decide to lend a hand. *Personality and Social Psychology Bulletin, 30*, 461–474.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 629–654.

Barsalou, L. W. (1992). *Cognitive psychology: An overview for cognitive scientists*. Hillsdale, NJ: Erlbaum.

Borkenau, P. (1990). Traits as ideal-based and goal-derived social categories. *Journal of Personality and Social Psychology*, *58*, 381–396.

Borkenau, P. (1991). Proximity to central tendency and usefulness in attaining goals as predictors of prototypicality for behaviour-descriptive categories. *European Journal of Personality*, *5*, 71–78

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review*, *90*, 105–126.

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

Fillmore, C. J. (1968). The case for case. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory* (pp. 1–88). New York: Holt, Rinehart & Winston.

Fleeson, W., Zirkel, S., & Smith, E. E. (1995). Mental representations of trait categories and their influences on person perception. *Social Cognition*, *13*, 365–397.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 89–150). New York: McGraw-Hill.

Graesser, A. C., & Clark, L. F. (1985). *Structures and procedures of implicit knowledge*. Norwood, NJ: Ablex.

Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT Press.

John, O. P. (1986). How shall a trait be called: A feature analysis of altruism. In A. Angleitner, A. Furnham, & G. van Heek (Eds.), *Personality psychology in Europe: Current trends and controversies* (pp. 117–140). Berwyn, PA: Swets North America.

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.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28*, 107–128.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits and behaviors: A parallel constraint satisfaction theory. *Psychological Review*, *103*, 284–308.

Miller, L. C., & Read, S. J. (1991). On the coherence of mental models of persons and relationships: A knowledge structure approach. In G. J. O. Fletcher & F. Fincham (Eds.), *Cognition in close relationships* (pp. 69–99). Hillsdale, NJ: Erlbaum.

Read, S. J. (1987). Constructing causal scenarios: A knowledge structure approach to causal reasoning. *Journal of Personality and Social Psychology, 52*, 288–302.

Read, S. J. (2004). *An integrative model of causal learning and causal reasoning using a feedback neural network*. Unpublished manuscript, University of Southern California, Los Angeles.

Read, S. J., Brownstein, A., & Miller, L. C. (2005). *Uncovering the components of traits*. Unpublished manuscript, University of California, Los Angeles.

Read, S. J., Jones, D. K., & Miller, L. C. (1990). Traits as goal-based categories: The importance of goals in the coherence of dispositional categories. *Journal of Personality and Social Psychology, 58,* 1048–1061.

Read, S. J., & Lalwani, N. (1998). *A narrative model of trait inferences: The stories traits tell.* Unpublished manuscript, University of Southern California, Los Angeles.

Read, S. J., & Lincer-Hill, H. (1998). *Explanatory coherence in dispositional inference.* Unpublished manuscript, University of Southern California, Los Angeles.

Read, S. J., & Marcus-Newhall, A. R. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology, 65*, 429–447.

Read, S. J., & Miller, L. C. (1989). Interpersonalism: Toward a goal-based theory of persons in relationships. In L. Pervin (Ed.), *Goal concepts in personality and social psychology.* Hillsdale, NJ: Erlbaum.

Read, S. J., & Miller, L. C. (1993). Rapist or "regular guy": Explanatory coherence in the construction of mental models of others. *Personality and Social Psychology Bulletin, 19*, 526–540.

Read, S. J., & Miller, L. C. (1995). Stories are fundamental to meaning and memory: For social creatures, could it be otherwise? In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Advances in social cognition. Vol. 8: Knowledge and memory: The real story* (pp. 139–152). Hillsdale, NJ: Erlbaum.

Read, S. J., & Miller, L. C. (1998). On the dynamic construction of meaning: An interactive activation and competition model of social perception. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and behavior* (pp. 27–68). Mahwah, NJ: Erlbaum.

Read, S. J., Vanman, E. J., & Miller, L. C. (1997). Connectionism, parallel constraint satisfaction processes and Gestalt principles: (Re)introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review, 1*, 26–53.

Reeder, G. D., Hesson-McInnis, M., Krohse, J. O., & Scialabba, E. A. (2001). Inferences about effort and ability. *Personality and Social Psychology Bulletin, 27*, 1225–1235.

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology, 83,* 789–803.

Reeder, G. D., Pryor, J. B., Wohl, M. J. A., & Griswell, M. L. (2004). *Egocentric motive attribution*. Unpublished manuscript.

Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional attribution: Multiple inferences about motive-related traits. *Journal of Personality and Social Psychology, 86*, 530–544.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*, 435–467.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

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

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.

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

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology*, *47*, 237–252.

This page intentionally left blank

PART III

Reading One's Own Mind, Reading Other Minds

This page intentionally left blank

9

Perspective Taking as the Royal Avenue to Empathy

JEAN DECETY

Why and how do we care about others? If we put ourselves into the mental shoes of another person, how closely do we really feel what he or she feels? What cognitive and neural mechanisms account for empathic understanding? Despite the obvious importance of this ability, it is a difficult concept to define (see Ickes, 1997). The aim of this chapter is to integrate several sources of knowledge, including developmental science, social psychology, and cognitive neuroscience, into a coherent model that can help to illuminate the basic mechanisms involved in human empathy. This model posits that empathy consists of both affective and cognitive components and that the capacity to adopt the perspective of the other is a key aspect of human empathy.

The central theme of this chapter is that one's own perspective is the default mode (and the prepotent one) by which we relate to others. We see others as similar to ourselves on a variety of dimensions and consequently assume that they act as we act, know what we know, and feel what we feel. This default mode is based on a shared representations mechanism between self and other (Decety & Sommerville, 2003) driven by the automatic link between perception and action (Jackson & Decety, 2004). However, for successful social interaction, and empathic understanding in particular, an adjustment must operate on these shared representations. While the projection of self-attributes onto the other does not necessitate any significant store of knowledge about the other, empathic understanding requires the inclusion of other characteristics within the self. In addition, a complete merging or confusion of self and other's feelings is not the goal of empathy. Hence, the mental flexibility to intentionally adopt the perspective of the other and self-awareness are two important components of empathy. One needs to calibrate one's own perspective that has been activated by the interaction with the other or even its mere imagination. Such calibration requires executive resources that are subserved by the prefrontal cortex, as demonstrated by neuroimaging studies in healthy participants as well as neuropsychological observations.

In this theoretical view, no one component (i.e., emotion sharing, mental flexibility/regulation, or self–other awareness) can separately account for the potential of human empathy (see Table 9.1). These components are intertwined and must interact with one another to produce the experience of empathy. For instance, sharing emotion without self-awareness corresponds to the phenomenon of emotional contagion, and is different from empathy.

PERSPECTIVE TAKING AS A SOURCE OF HUMAN EMPATHY

Empathy may be initiated by a variety of situations—for instance, when one sees another person in distress or discomfort, when one imagines someone else's behavior, when reading fiction, or when seeing a moving

TABLE 9.1. Basic Principles for a Functional Architecture of Human Empathy (see Decety & Jackson, 2004)

- This model of empathy combines both representational components (i.e., memories that are localized in distributed neural networks that encode information and, when activated, enable access to this stored information) and processes (i.e., computational procedures that are localized and are independent of the nature or modality of the stimulus that is being processed).
- These components are implemented in specific and dissociable neural systems.
- Like many emotion-related processes, the functioning of some components involved in empathy occurs implicitly, and sometimes without awareness (e.g., emotion sharing). Other components require explicit processing, such as taking the perspective of the other, representing our own thoughts and feelings as well as those of others, as well as some aspects of emotion regulation.

TV report. However, in these situations, empathy requires one to adopt more or less consciously the subjective point of view of the other. A more straightforward instance is when a psychotherapist attempts to adopt the mental world of her client (e.g., Reik, 1949).

Perspective taking is acknowledged as an important source of human empathy (Batson, 1991). An experiment by the social psychologist Ezra Stotland (1969) illustrates the effect of perspective taking on generating empathy. In his experiment, the participants watched someone else whose hand was strapped in a machine that they were told generated painful heat. One group of subjects was told just to watch the target person carefully, another group of subjects was asked to imagine the way the target was feeling, and one more group was told to imagine themselves in the target's place. Both physiological (i.e., palmar sweating and vasoconstriction) and verbal measures of empathy showed that the deliberate acts of imagination produced a greater response than just watching. In past decades, Batson has conducted a variety of studies that demonstrate the effectiveness of perspective-taking instructions in inducing empathy. An important aspect of Batson's theoretical framework is that empathy-inducing conditions do not compromise the distinction between the self and other (e.g., Batson, Early, & Salvarani, 1997). This view is congruent with Carl Rogers's client-centered therapy school, in which empathy means to perceive the internal frame of reference of another person with accuracy and with the emotional components and meanings that pertain thereto as if one were the person, but without losing the as-if condition (Rogers, 1959).

SELF-PERSPECTIVE AS THE DEFAULT MODE

The ability to take the psychological perspective of the other is considered an indispensable element in the fully developed mature theory of mind. Developmental research also indicates that perspective-taking ability develops gradually. In the affective domain, it is around 18 months that children demonstrate an emerging awareness of the subjectivity of other people's emotions. By that age, infants seem to understand, for instance, that they should give an experimenter a piece of food that the experimenter reacts to with apparent happiness (e.g., broccoli) rather than one toward which the experimenter acts disgusted (e.g., crackers), even when they themselves prefer the latter food; in contrast, 14-month-olds do not show this understanding (Repacholi & Gopnik, 1997). This finding appears to be the first empirical evidence that infants of this age have at least some limited ability to reason non-egocentrically about people's desires (Flavell, 1999). However, the mere fact that people have the ability to distinguish between their own perspective and the perspective of others does not mean that adults reliably and spontaneously use this ability when reasoning about them (Barr & Keysar, Chapter 17, this volume). Indeed, even adults frequently make a less sharp distinction between what they know, or believe they know, and what they assume others do. An elegant series of experiments conducted by Keysar, Lin, and Barr (2003) demonstrated that adult subjects show a tendency to infer that others have the same knowledge (and beliefs) as they do, even when they are aware that the others have a different point of view.

Several social and developmental psychologists have suggested, and documented through empirical work, that our default mode to reasoning about others is biased toward the self-perspective, and this is a general feature of human cognition (e.g., Nickerson, 1999). For instance, we have the tendency to believe that our actions and appearance are more likely to be noticed, judged, and remembered by others than is actually the case (Gilovich, Kruger, & Medvec, 2002). We are inclined to impute our own knowledge to others and overestimate what they know. Recent research also indicates that people's predictions about the feelings of others who are in a situation that arouses drive states (i.e., motivations caused by bodily needs such as exhaustion, hunger, and thirst) are based largely on their predictions of how they themselves would feel in that situation (see Van Boven & Loewenstein, Chapter 18, this volume).

FROM PERCEPTION–ACTION COUPLING TO EMOTION SHARING

The view that self-perspective is the default way that we socially operate is coherent with the perception–action coupling mechanism. This notion reflects the idea that the perception of a given behavior in another individual automatically activates one's own motor representations of that behavior (Preston & de Waal, 2002). Such a view is grounded in the fundamental physiological properties of the nervous system regarding the continuity between action and cognition, which is primarily based on perception and action cycles (Sperry, 1952). These processes are functionally intertwined: perception is a means to action, and action is a means to perception, and both operate right after birth. The automatic mapping between self and other is also supported by considerable empirical literature in the domain of perception and action in both cognitive and social psychology, which has been marshaled under the common-coding theory (Dijksterhuis & Bargh, 2001; Prinz, 1997). This theory claims parity between perception and action. Its core assumption is that actions are coded in terms of the perceivable effects (i.e., the distal perceptual events) they should generate (e.g., Knoblich & Flach, 2003). This theory also states that perception of an action should activate action representations to the degree that the perceived and the represented action are similar. Furthermore, when two individuals socially interact with each other, this overlap creates shared representations, that is, neural networks that are temporarily and simultaneously activated in both agents' brains (Decety & Chaminade, 2003a; Decety & Sommerville, 2003; Jeannerod, 1999).

In neuroscience, evidence for this action–perception coupling ranges from electrophysiological recordings in monkeys in which mirror neurons that fire both during goal-directed actions and observation of actions performed by another individual to functional neuroimaging experiments in humans (Blakemore & Decety, 2001, for a review). The combined results of these studies demonstrate that when individuals observe the actions performed by others, they use the same neural mechanisms as when they produce the action themselves. Moreover, a number of neuroimaging studies have shown that similar brain areas including the premotor and posterior parietal cortex are activated during imagining one's own action, imagining another's action, and imitating actions performed by a model (Decety & Chaminade, 2003b). Such a coupling mechanism thus offers an interesting foundation for intersubjectivity because it provides a functional bridge between first-person information and third-person information.

This mechanism accounts (at least partly) for emotion processing and empathy, as suggested by Preston and de Waal (2002). In this model, perception of emotion activates the neural mechanisms that are responsible for the generation of emotions (Adolphs, 2002). Such a system prompts the observer to resonate with the state of another individual. From the study of a large group of neurological patients, Adolphs, Damasio, and Tranel (2002) conjectured that the reconstruction of knowledge about emotions expressed by other people relies on a covert simulation of how the emotion would feel in the perceiver.

Neuroimaging studies have contributed evidence for the coupling mechanism of emotion processing. For instance, Ekman and Davidson (1993) were able to demonstrate similar patterns of electroencephalographic activity for spontaneous and voluntary forms of smiling. Recently, a functional magnetic resonance imaging (fMRI) experiment confirmed and extended these findings by showing that when participants are required to observe or to imitate facial expressions of various emotions, increased hemodynamic activity is detected in the superior temporal sulcus, the anterior insula and the amygdala, as well as areas of the premotor cortex corresponding to the facial representation (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003).

Even though this coupling mechanism is necessary for emotion sharing (and certainly emotion contagion), it is not sufficient for empathic understanding. Moreover, the overlap between self-related and other-related neural processing is not complete. There are specific subcircuits within the premotor, prefrontal, and parietal cortices that handle information either from the self or the other (e.g., Ruby & Decety, 2001; Seger, Stone, & Keenan, 2004). A recent fMRI experiment by Ramnani and Miall (2004) has shown that the motor system is engaged when participants use arbitrary visual cues to prepare their own actions and also when they use the same cues to predict the actions of other people. However, these two tasks activate separate subcircuits within the premotor cortex.

MENTAL FLEXIBILITY TO ADOPT THE PERSPECTIVE OF THE OTHER

For successful social interaction and empathic understanding in particular, an adjustment must operate on the representations elicited by the perception–action coupling mechanism. One needs to regulate one's own perspective that has been activated by the interaction with the other. Such regulation is important in modulating one's own vicarious emotion so that it is not experienced as aversive. Previous research has shown that emotion regulation is positively related to feelings of concern for the other person (e.g., Derryberry & Rothbart, 1988). In contrast, people who experience their emotions intensely, especially negative emotions, are prone to personal distress—that is, an aversive emotional reaction, such as anxiety or discomfort, based on the recognition of another's emotional state or condition (Eisenberg, Shea, Carlo, & Knight, 1991). Furthermore, a complete merging or confusion of self and other is not the goal of empathy (Batson, 1987, 1991; Decety & Jackson, 2004; Hodges & Wegner, 1997; Ickes, 2003). Indeed, an essential aspect of empathy is to recognize the other person as like the self while maintaining a clear separation between self and other. Hence, mental flexibility and self-regulation are important components of empathy.

The flexibility and self-regulation involved in empathy require executive function resources such as executive inhibition (i.e., the deliberate suppression of a cognition or response to achieve an internally represented goal). This inhibitory component is needed to regulate and tone down the prepotent (default) self-perspective and to allow the evaluation of the other's perspective. It is also congruent with the empathy–altruism hypothesis, which claims that a distinction between self and other is required rather than a merging between them (Batson, 1991).

FUNCTIONAL NEUROIMAGING OF PERSPECTIVE TAKING

A series of three neuroimaging studies performed by my research team investigated in healthy volunteers the neural underpinning of perspective taking in three different modalities (i.e., motoric, conceptual, and emotional) of self–other representations. In a first study, participants were scanned while they were asked to either imagine themselves performing a variety of everyday actions (e.g., winding a watch up) or imaging the experimenter doing similar actions (Ruby & Decety, 2001). Both conditions were associated with common activation in the supplementary motor area, premotor cortex, and the occipitotemporal region. In addition, taking the perspective of the other to mentally simulate his or her behavior resulted in selective activation of the frontopolar cortex and right inferior parietal lobule.

In a second study, medical students were shown a series of affirmative health-related sentences (e.g., taking antibiotic drugs causes general fatigue) and were asked to judge their truthfulness either according to their own perspective (i.e., as experts in medical knowledge) or according to the perspective of a layperson (Ruby & Decety, 2003). The set of activated regions recruited when the participants put themselves in the shoes of a layperson to simulate his or her knowledge included the medial prefrontal cortex, the frontopolar, the posterior cingulate cortex and right inferior parietal lobule.

In a third study, the participants were presented with short written sentences that depicted real-life situations (e.g., someone opens the toilet door that you have forgotten to lock) that are likely to induce social emotions (e.g., shame, guilt, pride) or other situations that are emotionally neutral (Ruby & Decety, 2004). They were asked to imagine how they would feel if they were in those situations and how their mother would feel in those situations. The mother was chosen as the target of empathy because she was the participants' best-known person. Hemodynamic activation was detected in the frontopolar cortex, the ventromedial prefrontal cortex, the medial prefrontal cortex, the posterior cingulate cortex, and the right inferior parietal lobule when the participants adopted the perspective of their mother, regardless of the affective content of the depicted situations. Cortical regions that are involved in emotional processing were found activated in the conditions that integrated emotion-laden situations, including the amygdala and the temporal poles. The amygdala is acknowledged to be critical for normal judgments about the internal states of others (Adolphs, 2002). It is thus interesting to detect its activation for both self- and other-imagined emotional reactions.

THE SENSE OF AGENCY

The ability to recognize oneself as the agent of an action, thought, or desire (the sense of agency) is crucial for attributing a behavior to its proper agent. Thus, the distinction between self-generated actions and actions produced by others is a key function for self-recognition. Such a tracking or monitoring mechanism is necessary for an empathic response to take place. Without such a mechanism, confusion between self and other would occur. A striking finding of the aforementioned neuroimaging studies of perspective taking is the systematic activation of the right inferior parietal cortex when participants adopt the perspective of another. Recent research indicates that the right inferior parietal cortex in conjunction with prefrontal areas may be critical in distinguishing the self from the other, and therefore navigating the shared representations (Jackson & Decety, 2004). The inferior parietal cortex is a heteromodal association area that receives input from the lateral and posterior thalamus, as well as visual, auditory, somesthetic, and limbic areas. It has reciprocal connections to the prefrontal cortex and to the temporal lobes (Eidelberg & Galaburda, 1984). These multiple connections confer on this region a role in the elaboration of an image of the body in space and time (Benton & Silvan, 1993), on which the sense of agency depends. Accumulated empirical evidence indicates that the inferior parietal cortex in conjunction with the prefrontal cortex plays a major role in the sense of agency—in distinguishing between self-produced actions and actions generated by others (see Blakemore & Frith, 2003, for a review). For instance, Farrer and Frith (2002) scanned individuals engaged in watching a moving dot on a computer screen. In some trials, the participants were in control of the dot's movements, whereas in other trials someone else controlled the dot. They found increased activity in the right inferior parietal cortex when the dot was controlled by the other, and increased activity in the anterior insula when the dot was controlled by the self. Similarly, activation in the right inferior parietal lobe was found in reciprocal imitation when participants were aware (and observed) that their actions were being imitated online by another person (Decety, Chaminade, Grèzes, & Meltzoff, 2002).

CONTRIBUTIONS FROM NEUROLOGICAL STUDIES

Many neuropsychiatric disorders are associated with empathy deficits (e.g., Asperger's syndrome, psychopathy, and stroke). This is consistent with a model of empathy that involves parallel and distributed processing in a number of dissociable computational mechanisms (Decety & Jackson, 2004). Empathic processing may be impaired after focal lesions of the prefrontal cortex (Eslinger, 1998). Patients with bilateral lesions of the orbitofrontal cortex were found to be impaired in the "faux pas" task (Stone, Baron-Cohen, & Knight, 1998), which requires both an understanding of false belief and an appreciation of the emotional impact of a statement on the listener. A study conducted by Stuss, Gallup, and Alexander (2001) extended this finding by showing that only lesions in the right orbitofrontal produce such a deficit. In addition, several other patient studies reported a relationship between the deficit in empathy and performance of cognitive flexibility tasks among patients with lesions in the dorsolateral cortex, while those with orbitofrontal cortex lesions were more impaired in empathy but not in cognitive flexibility (Grattan, Bloomer, Archambault, & Eslinger, 1994). Furthermore, a study by Shamay-Tsoory, Tomer, Berger, and Aharon-Peretz (2003) reported that among patients with posterior lesions only those with damage to the right hemisphere (parietal cortex) were impaired in empathy. Another recent study by the same group tested patients with lesions of the ventromedial prefrontal cortex or dorsolateral prefrontal cortex with three theory-of-mind tasks (second-order beliefs and faux pas) differing in the level of emotional processing involved (Shamay-Tsoory, Tomer, Berger, Goldsher, & Aharon-Peretz, in press). The authors found that patients with ventromedial lesions were most impaired in the faux-pas task but presented normal performance in the second-order belief tasks. They further argued that in order to detect a faux pas one is required not only to understand the knowledge of the other but also to have empathic understanding of his or her feelings.

These studies suggest that different parts of the right prefrontal cortex are involved in the capacity to reason about the feelings of others, including the ability to adopt the perspective of others. However, it is not yet clear what specific processes each subregion subserves. For instance, one case study of a patient with orbitofrontal damage showed impaired real-life social cognition deficits despite intact neuropsychological performance (Cicerone & Tanenbaum, 1997). The patient, who suffered from traumatic orbitomedial frontal lobe damage, demonstrated good neurocognitive recovery but a lasting profound disturbance of emotional regulation and social cognition. Indeed, while measures associated with frontal lobe functions were found normal, the patient remained impaired on tasks requiring the interpretation of social situations, which mirrored her impairment in real-life functioning. The role of the orbitofrontal cortex in emotion regulation is further supported by a study involving five cases with similar orbitofrontal lesions (Beer, Heerey, Keltner, Scabini, & Knight, 2003). Following comparison with healthy individuals on a number of social/emotional measures, this study suggests that deficient behavioral regulation is associated with inappropriate self-conscious emotions, or faulty appraisals, that reinforce maladaptive behavior. Moreover, the authors provided evidence that deficient behavioral regulation is associated with impairments in interpreting the self-conscious emotions of others.

The study of degenerative neurological diseases has also supplied evidence for relatively distinct routes to social cognition and empathy deficits. For instance, Snowden and colleagues (2003) have shown that both patients with a frontotemporal dementia (FTD, a predominantly neocortical disorder associated with deficits in frontal executive functions) as well as patients with Huntington's disease (HD, a predominantly subcortical disorder characterized by involuntary movements) present difficulties in tasks of social cognition. However, the two patient groups display qualitatively different patterns of results, which suggest that the deficits of patients with FTD may be attributed to a breakdown in theory of mind while those of patients with HD appear to be associated with faulty inferences drawn from social situations. Interestingly, patients with both HD and FTD lack sympathy and empathy, but for different reasons. In the former group, the loss of empathy arises more at an emotional than a cognitive level, while patients with FTD live in an egocentric world in which they do not ascribe independent mental states to others. An interesting finding from a voxel-based morphometry analysis (i.e., a technique used for detection of regional brain atrophy) on patients with FTD revealed that atrophy in bilateral temporal lobe and medial orbitofrontal structures correlated with loss of cognitive empathy and that atrophy to the temporal pole correlated significantly with loss of emotional empathy (Rankin, Gorno-Tempini, Weiner, & Miller, 2003). This result is consistent with distinct neural underpinnings for the cognitive and affective aspects of empathy.

The amygdala is another region critically involved in social cognition that seems to play a role in empathy (Ruby & Decety, 2004). Its lesion can produce deficits in both mental state reasoning and empathy, as demonstrated by Stone and colleagues (Stone, Baron-Cohen, Calder, Keane, & Young, 2003), who found right-sided lesions associated with affective state attributions. The role of the amygdala in mental state attribution was further investigated in a large group of neurological patients by Shaw and colleagues (2004). They found that such a deficit was observed in individuals who had amygdala damage during childhood but not when lesions occurred later. This shows that theory of mind and affective processing are partly dissociable.

In sum, the neuropsychological evidence suggests that different acquired pathologies can lead to reduced empathic ability, which can be traced to deficits in distinct neural systems.

CONCLUSION: EMPATHY AS A COMPLEX SOCIAL BEHAVIOR

I have argued that, in addition to emotional sharing and self-awareness, a key aspect of human empathy is the ability to consciously adopt the perspective of the other. This process is essential for adjusting and regulating the self-perspective that is automatically triggered in social interaction and responsible for unconscious projective phenomena. It involves setting aside one's own current perspective, attributing a mental state to the other person, and then inferring the likely content of the mental state of that person (Leslie, 1987). Forming an explicit representation of another person's feelings thus necessitates an additional mechanism beyond the shared representation mechanism (see Decety & Jackson, 2004; Ickes, 1997). It requires that second-order representations of the other be available to consciousness (a decoupling mechanism between first-person and third-person information), for which the anterior paracingulate cortex seems to play a unique function (Frith & Frith, 2003). Thus, in the view developed here, empathy is not a simple resonance of affect between the self and the other, and it is perspective taking that creates an explicit representation of the other. This makes empathy as described here a representational capacity. Recent neuroimaging investigations of empathy for pain in others, which stressed the role of the anterior cingulate cortex and insula, support such a view (Jackson, Meltzoff, & Decety, 2004; Singer et al., 2004).

There is no unitary empathy system (or module) in the brain. Rather, there are multiple dissociable systems involved in the experience of empathy. Shared representations refer to distributed patterns of neural activation in two individuals who socially interact. These patterns, temporarily activated, are widely distributed in the brain, and their location varies according to the processing domain, the particular emotion, and the stored information. Moreover, cognitive processes that exert a top-down control on these shared representations are mediated by specific subregions of the prefrontal cortex, namely, the frontopolar and ventromedial cortex and anterior paracingulate/medial areas. Another process associated with empathy is the sense of agency, for which the right inferior parietal lobule in conjunction with the prefrontal cortex plays a pivotal role. Each subsystem may be selectively damaged, which may lead to specific neuropsychological disorders.

Finally, empathy in the present model is a motivated behavior that more often than commonly believed is triggered voluntarily. This makes empathy a psychological capacity prone to social-cognitive intervention such as through training or enhancement programs for the sake of various goals (e.g., reeducation of antisocial personalities; early consultation with at-risk children; and training of psychotherapists or physicians).

ACKNOWLEDGMENTS

Sara D. Hodges, Bertram F. Malle, Daniel Bernstein, and an anonymous reviewer provided valuable comments on an earlier version of this chapter.

REFERENCES

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews, 1*, 21–62.

Adolphs, R., Damasio, H., & Tranel, D. (2002). Neural systems for recognition of emotional prosody: A 3–D lesion study. *Emotion, 2*, 23–51.

Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic? In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 20, pp. 65–122). New York: Academic Press.

Batson, C. D. (1991). Empathic joy and the empathy–altruism hypothesis. *Journal of Personality and Social Psychology, 61*, 413–426.

Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective taking: Imagining how another feels versus imagining how you would feel. *Personality and Social Psychology Bulletin, 23*, 751–758.

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.

Benton, A., & Silvan, A. B. (1993). Disturbance of body schema. In K. M. Heilman & E. Valenstein (Eds.), *Clinical neuropsychology* (pp. 123–140). Oxford, UK: Oxford University Press.

Blakemore, S.-J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2*, 561–567.

Blakemore, S.-J., & Frith, C. D. (2003). Self-awareness and action. *Current Opinion in Neurobiology, 13*, 219–224.

Carr, L., Iacoboni, M., Dubeau, C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. *Proceedings of National Academy of Sciences, 100*, 5497–5502.

Cicerone, K. D., & Tanenbaum, L. N. (1997). Disturbance of social cognition after traumatic orbitofrontal brain injury. *Archives of Clinical Neuropsychology, 12*, 173–188.

Decety, J., & Chaminade, T. (2003a). Neural correlates of feeling sympathy. *Neuropsychologia, 41*, 127–138.

Decety, J., & Chaminade, T. (2003b). When the self represents the other: A new cognitive neuroscience view of psychological identification. *Consciousness and Cognition, 12*, 577–596.

Decety, J., Chaminade, T., Grèzes, J., & Meltzoff, A. N. (2002). A PET exploration of the neural mechanisms involved in reciprocal imitation. *NeuroImage, 15*, 265–272.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews, 3*, 71–100.

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and others: A social cognitive neuroscience view. *Trends in Cognitive Science, 7*, 527–533.

Derryberry, D., & Rothbart, M. K. (1988). Arousal, affect, and attention as components of temperament. *Journal of Personality and Social Psychology, 55*, 958–966.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception–behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology, 33*, 1–40.

Eidelberg, D., & Galaburda, A. M. (1984). Inferior parietal lobule. *Archives of Neurology, 41*, 843–852.

Eisenberg, N., Shea, C. L., Carlo, G., & Knight, G. (1991). Empathy related responding and cognition: A "chicken and the egg" dilemma. In W. Kurtines & J. Gewirtz (Eds.), *Handbook of moral behavior and development: Vol. 2. Research* (pp. 63–68). Hillsdale, NJ: Erlbaum.

Ekman, P., & Davidson, R. J. (1993). Voluntary smiling changes regional brain activity. *Psychological Science, 4,* 342–345.

Eslinger, P. J. (1998). Neurological and neuropsychological bases of empathy. *European Neurology, 39*, 193–199.

Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs. another person as being the cause of an action: The neural correlates of the experience of agency. *NeuroImage, 15*, 596–603.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology, 50*, 21–45.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalazing. *Philosophical Transactions of the Royal Society, London B, 358*, 459–473.

Gilovich, T., Kruger, J., & Medvec, V. H. (2002). The spotlight effect revisited: Overestimating the manifest variability of our actions and appearance. *Journal of Experimental Social Psychology, 38*, 93–99.

Grattan, L. M., Bloomer, R. H., Archambault, F. X., & Eslinger, P. J. (1994). Cognitive flexibility and empathy after frontal lobe lesion. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 7*, 251–257.

Hodges, S., & Wegner, D. M. (1997). Automatic and controlled empathy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 311–339). New York: Guilford Press.

Ickes, W. (1997). *Empathic accuracy*. New York: Guilford Press.

Ickes, W. (2003). *Everyday mind reading*. New York: Prometheus Books.

Jackson, P. L., & Decety, J. (2004). Motor cognition: A new paradigm to study self-other interactions. *Current Opinion in Neurobiology, 14*, 259–263.

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, 23*, 744–751.

Jeannerod, M. (1999). To act or not to act: Perspectives on the representation of actions. *Quarterly Journal of Experimintal Psychology, 52A*, 1–29.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind in adults. *Cognition, 89*, 25–41.

Knoblich, G., & Flach, R. (2003). Action identity: Evidence from self-recognition, prediction, and coordination. *Consciousness and Cognition, 12*, 620–632.

Leslie, A. M. (1987). Pretence and representation: The origins of "theory of mind." *Psychological Review, 94*, 412–426.

Nickerson, R. S. (1999). How we know and sometimes misjudge what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 126*, 737–759.

Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences, 25*, 1–72.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*, 129–154.

Ramnani, N., & Miall, R. (2004). A system in the human brain for predicting the actions of others? *Nature Neuroscience, 7*, 85–90.

Rankin, K. P., Gorno-Tempini, M. L., Weiner, M. W., & Miller, B. L. (2003). *Neuroanatomy of impaired empathy in frontotemporal dementia*. Paper presented at the 55th annual meeting of the American Academy of Neurology, Honolulu.

Reik, T. (1949). *Character analysis*. New York: Farrar, Strauss & Giroux.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- to 18-month-olds. *Developmental Psychology, 33*, 12–21.

Rogers, C. (1959). A theory of therapy, and interpersonal relationships as developed in the client-centered framework. In J. S. Koch (Ed.), *Psychology: A study of a science: Vol. 3. Formulations of the person in the social context* (pp. 184–256). New York: McGraw-Hill.

Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: A PET investigation of agency. *Nature Neuroscience, 4*, 546–550.

Ruby, P., & Decety, J. (2003). What you believe versus what you think they believe: A neuroimaging study of conceptual perspective taking. *European Journal of Neuroscience, 17*, 2475–2480.

Ruby, P., & Decety, J. (2004). How would you feel versus how do you think she would feel? A neuroimaging study of perspective taking with social emotions. *Journal of Cognitive Neuroscience, 16*, 888–899.

Seger, C. A., Stone, M., & Keenan, J. P. (2004). Cortical activations during judgments about the self and another person. *Neuropsychologia, 42*, 614–629.

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., & Aharon-Peretz, J. (2003). Characterization of empathy deficits following prefrontal brain damage: The role of right ventromedial prefrontal cortex. *Journal of Cognitive Neuroscience, 15*, 1–14.

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., Goldsher, D., & Aharon-Peretz, J. (in press). Impaired affective theory of mind is associated with ventromedial prefrontal damage. *Cognitive and Behavioral Neurology.*

Shaw, P., Lawrence, E. J., Radbourne, C., Bramham, J., Polkey, C. E., & David, A. S. (2004). The impact of early and late damage to the human amygdala on "theory of mind" reasoning. *Brain, 127*, 1535–1548.

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*, 1157–1161.

Snowden, J. S., Gibbons, Z. C., Blackshaw, A., Doubleday, E., Thompson, J., Craufurd, D., et al. (2003). Social cognition in frontotemporal dementia and Huntington's disease. *Neuropsychologia, 41*, 688–701.

Sperry, R. W. (1952). Neurology and the mind-body problem. *American Scientist, 40*, 291–312.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience, 10*, 640–646.

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

Stotland, E. (1969). Exploratory Investigations of Empathy. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 4, pp. 271–314). New York: Academic Press.

Stuss, D. T., Gallup, G., & Alexander, M. P. (2001). The frontal lobes are necessary for theory of mind. *Brain, 124*, 279–286.

10

Everyday Solutions to the Problem of Other Minds

Which Tools Are Used When?

DANIEL R. AMES

Intuiting what the people around us think, want, and feel is essential to much of social life. To bargain, we make assumptions about what a partner prefers or wants to avoid. To persuade, we try to intuit an audience's beliefs. To console, we infer the depth of a friend's grief. Whether we are sizing someone up or seducing him or her, assigning blame or extending our trust, we are very nearly always performing the ordinary magic of mindreading. In some sense, of course, we cannot read minds. Some scholars have gone so far as to declare the "problem of other minds"—whether a person can know if anyone else has thoughts and, if so, what they are—intractable. And yet countless times a day, we solve such problems with ease, if not perfectly then at least to our own satisfaction. What strategies underlie these everyday solutions? And how are these tools employed?

The perceiver's bag of tricks has revealed much of itself to scholars, and there is no need to invoke magic on the mindreader's behalf (see Figure 10.1). Perceivers rely in part on evidence-based induction, working with the raw data of local physical behavior (a grasping hand that

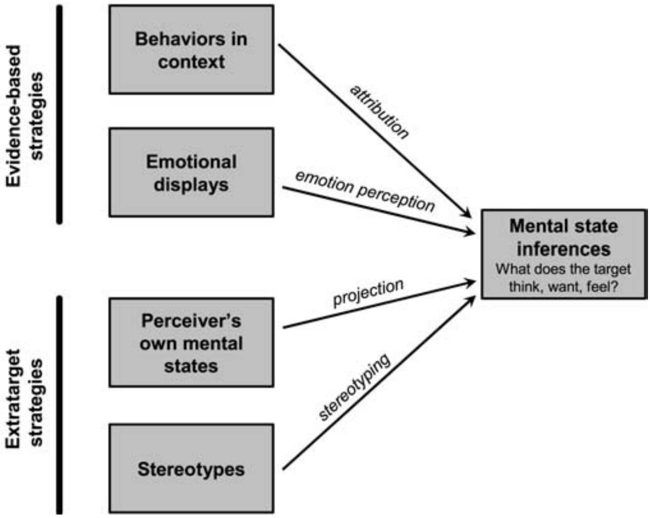

FIGURE 10.1. Mindreading strategies.

"wants") as well as more general chains of action in context (the colleague whose generosity to superiors but not subordinates signals ingratiation). Perceivers also take account of nonverbal behavior and displays of emotion, not only to intuit a target's fleeting feelings but also as a window onto the target's underlying goals. In short, perceivers readily work from the visible evidence of human behavior to posit invisible underlying mental states.

Along with these inductive approaches, perceivers often work from information and premises that are exogenous to a target and his or her behavior. Perceivers may consult their own general and enduring mental states and ascribe them to others ("I love cats and assume you do, too") and also engage in more effortful and situation-specific forms of perspective taking ("I'd be embarrassed if I were in your shoes"). Perceivers also invoke stereotypes, assuming a target's mental states correspond to those likely to be held by some larger category of people ("Jocks hate romantic comedies"). Thus, even without immediate evidence of a target's behavior, a perceiver may reason from stereotypes or from his or her own self to intuit what the target is thinking, wanting, or feeling.

There is every reason to think that *all* of these tools, as well as some others I do not elaborate here, play some role in everyday mindreading.[1] Sometimes multiple tools may be used simultaneously or in succession. Yet surely we rely on some tools at some times more than others. The central question I wish to address here, and that I suggest is not yet well answered, is "Which tool is used when?" More complete and elegant answers to this question await us, but in the meantime I develop several claims in this chapter about which-tool-when contingencies. For instance, I suggest that perceptions of general similarity guide a trade-off between projection (ascribing one's own beliefs and desires to others) and stereotyping. I also claim that displays of self-conscious emotions, such as shame, can override inferences of malicious intent from harmful behavior, though only in the short run. Some of these claims are backed with mounting evidence, while others are more speculative. My hope, beyond aspiring to be mostly right, is that this discussion will provoke others to address the question of "Which tool when?" To the extent that mindreading governs much of social life, scholars are well served to understand how perceivers pursue it. And to the extent that perceivers pursue it with multiple inferential strategies, scholars are well served to explore the conditions governing when different approaches are employed.

EVIDENCE-BASED STRATEGIES: BEHAVIOR AND AFFECTIVE DISPLAYS

From our earliest days, we are voracious and deft observers of ordinary human behavior. This consumption of overt acts is often in the service of intuiting unseen mental states. Work by Meltzoff and colleagues (e.g., Meltzoff & Brooks, 2001), for instance, shows that by even 6 months of age infants know that a grabbing hand entails *wanting*. When adults view an impoverished cartoon in which animated dots represent a person's arm and hand knocking on a door, they can readily intuit the phantom knocker's fear, excitement, and happiness (Pollick, Paterson, Bruderlin, & Sanford, 2001). From the smallest arch of a teasing eyebrow to the profound slouch of a disinterested onlooker, we fluidly intuit what others think from even the smallest scraps of static and dynamic evidence.

Scholars have also observed that perceivers are attuned to more elaborate and prolonged arcs of behavior as cues to underlying intentions. Social psychologists have long recognized that perceivers intuit an actor's intentions on the basis of behavior in the face of situational forces (e.g., Heider, 1958; Kelley, 1967) and from the pattern of options he or she chooses and forgoes (Jones & Davis, 1965; Newtson, 1974; see also Malle, 2004, for an overview).

Recent work has brought new attention to the role of situations in intuiting motives. Reeder and colleagues have shown that the same behaviors can imply vastly different motives depending on context (e.g., Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002; see also Reeder & Trafimow, Chapter 7, this volume). For instance, when the same act of aggression is performed in a situation where it is provoked rather than one where it is self-serving, perceivers reach substantially different inferences about the aggressor's motives and morality. Meanwhile, Kammrath, Mendoza-Denton, and Mischel (in press) have highlighted that perceivers attend to profiles of behaviors across multiple situations to disambiguate underlying motives, showing that mental state inferences flow not just from behavioral base rates (e.g., how frequently someone is pleasant) but also from patterns across contexts (e.g., frequency of being pleasant to authorities vs. peers).

As targets of other people's perceptions, people are more than actors—they are reactors. Emotional displays are often behaviors *about* behaviors: a person beams when proud of her work, blushes when embarrassed by her acts, and scowls when foiled in her pursuits. In this sense, people's emotional displays are not so much intentional goal-driven behavior as unintentional reactions that stem from the match between one's goals and one's own behavior and outcomes. For the mindreader perceiving a target, these displays can reveal much, not only about passing emotional states but also about these underlying enduring motives and goals.

In recent work, my colleagues and I (Ames, Johar, & Kammrath, 2004) have explored how a target's affective displays shape inferences not only about his or her emotions but also about his or her broader intentions. Our research explored how a target's displays of self-evaluative emotions, such as embarrassment and guilt, lead perceivers to ascribe intentions to the target that were at odds with the target's overt behaviors. Extending self-presentational ideas from Goffman (1959), we argued that emotional displays help perceivers intuit how the actor regards his or her behavior and outcomes: certain negative displays can signal "that's not me" (discounting), while positive ones can signal "that's definitely me" (augmenting). Our results show that displays of self-evaluative negative affect, such as shame, after harm lead to more positive impressions of harm-doers, compared to neutral or positive affect displays, for targets ranging from violent criminals to colleagues who declined to provide a favor (Ames et al., 2004). Conversely, displays of positive affect (vs. negative affect) while helping lead to more positive impressions of helpers. Both the discounting and augmenting effects on general impressions were mediated by inferred mental states: remorseful harm-doers and cheerful helpers were both seen as having more prosocial underlying motives.

But can these affective displays endlessly correct for engaging in harmful behaviors? Certainly not. Perceivers appear to be reasonably forgiving in isolated cases, recognizing that actions and outcomes do not necessarily correspond to a target's underlying intentions. In effect, an immediate and heartfelt apology or even a pained nonverbal expression from a new acquaintance is seen as more indicative of his innocent intentions than the red wine he has just spilled on our rug. Indeed, we may even like a genuinely and appropriately remorseful spiller (diagnosis: benign) more than a harmless guest (diagnosis: uncertain). But at some point repeated harm takes over, and the stain, as it were, becomes impossible to remove. This yields a first contingency:

> *Affect qualifies behavior in the near term: perceived remorseful affect can lead to ascriptions of good intent to harm-doers in the short run, but repeated harm drives long-run ascriptions of bad intent.*

In one study, we asked participants to imagine being paired with a classmate on a school project (Ames et al., 2004). Participants then read about a series of negative episodes involving the classmate (e.g., she showed up late for a meeting), recording their mental state inferences and impressions after each one. For some participants, the episodes were matched with photos of the classmate displaying negative self-conscious emotions, such as embarrassment and shame. For others, the photos were neutral. After the initial episode, participants seeing the negative affect displays had far more positive impressions of the target and generally assumed positive intentions (for instance, agreeing that the target wanted to be kind and respectful). After three episodes of harm (e.g., breaking the participant's cell phone, failing to complete work), impressions and mental state inferences were increasingly negative in both display conditions, though they were still significantly more positive in the remorse display condition. However, by six negative episodes (e.g., forgetting the participant's name, delivering bad work), the gap had closed: regardless of affective display, all participants had strongly negative assumptions about the target's mental states and personality. The repeated acts of harm outweighed the repeated displays of self-conscious emotions.

In sum, mindreaders often work from a target's behavior, considering not only local action (e.g., an act of help) but also arcs of behavior over time and across situations (e.g., helping authorities but not peers). Mindreaders also look to displays of emotions to intuit how actors feel about their own behavior and outcomes. In the short run, these affective displays may augment or discount behaviors; but in the longer run, behaviors appear to speak more loudly than affect.

EXTRATARGET STRATEGIES: PROJECTION AND STEREOTYPING

The wheels of inference rarely stop when bottom-up data are lacking. Indeed, in some cases, judgments made before any behavioral evidence is available may be hard to overturn despite an abundance of evidence to the contrary. In some of my own work on mindreading in two-party negotiations, I've found that the best predictor of a negotiator's post-interaction judgments about his or her partner's motives (e.g., "He wanted what was best for both of us") were not his or her partner's behaviors during the negotiation but the former's own expectations about the unfamiliar partner before the interaction even began.

I wish to focus on two sources of mindreading that are not based on behavioral evidence: *projection*, when the perceiver assumes a target has the same mental states that he or she has or would have, and *stereotyping*, when the perceiver assumes a target has mental states that correspond to some prior expectancy about a social category. In social psychology, projection has often been studied in connection with false consensus, a pattern of judgment that emerges when perceivers overestimate the extent to which others share their own attitudes, qualities, and behaviors. In most false consensus research, participants indicate their own yes/no response to a question and then estimate what percentage of other people would say yes/no. For instance, participants in a study by Krueger and Clement (1994) indicated agreement with the statement "I like to read newspaper articles on crime." Those who agreed, on average, expected 59% of people to also agree, while those who disagreed expected only 43% of people, on average, to agree with the statement. Most false consensus scholars interpret such a pattern of results as evidence of projection (e.g., Krueger, 1998). While various explanations for false consensus have been proposed, ranging from self-serving motivations to egocentric perception, the effect is sufficiently widespread that many researchers regard various forms of projection as primary, if not sovereign, forces in social judgment (see, e.g., Krueger, 2003; Van Boven & Loewenstein, in press).

Although projection has received lavish scholarly attention as a source of mindreading, stereotyping has been almost entirely ignored. The lack of integration appears on both ends: most stereotyping scholars examine the impact of stereotypes on broad impressions rather than mental state inferences, whereas most scholars of folk psychology have not given stereotypes an explicit role in everyday mindreading. Nonetheless, there is every reason to expect that the widespread importance of stereotyping in social judgment extends to mindreading. Scattered evidence confirms as much: beliefs about a target's race appear to have substantial effects on perceivers' expectations about what the target is feeling (e.g., Hugenberg & Bodenhausen, 2003) and trying to do (e.g., Sagar & Schofield, 1980).

Thus, projection and stereotyping each play some role in mindreading. But what governs when each will be invoked? And do they often function as alternative inferential strategies? Few answers to these questions exist, because most social psychological models focus on only one route or the other. Projection is rarely cast as an alternative to stereotyping, and, likewise, stereotyping is rarely described as an alternative to projection. Yet, I suggest that these judgmental strategies not only guide mindreading but also often displace each other in inferences. This yields a second contingency:

> *Perceived similarity governs projection and stereotyping: perceptions of general similarity to a target typically draw a mindreader toward projection and away from stereotyping; perceived dissimilarity does the opposite.*

I suggest that a greater sense of general similarity to a target evokes higher levels of projection of specific and novel mental states, whereas a diminished sense of general similarity to a target evokes higher levels of stereotyping. Importantly, I have repeatedly found that perceived similarity is often only weakly related to measures of actual similarity (see, e.g., Ames, 2004b). In my research, I have capitalized on the perceiver's tendency to overgeneralize similarity: when a perceiver finds one thing in common with a partner, he or she seems readily to expect that he or she has most everything (including a wide range of beliefs, desires, and feelings) in common. Upon learning that my new acquaintance shares my love for disco music, for instance, I may willingly intuit that she also shares my libertarian values and would be just as embarrassed as I would be if asked to speak in public.

This effect emerges for dissimilarities as well: upon learning she differs from a target in some particular way, the perceiver may assume she differs in many ways. When such a gulf opens between the self and other, stereotypes are likely to rush in. Thus, when I learn that my new Canadian collaborator hates my favorite comedian, I may feel a chasm between us and turn to my national stereotypes—however baseless—intuiting that, unlike me, he loves playing hockey and eating back bacon.

Evidence for these effects is accumulating. One set of studies (Ames, 2004a) manipulated perceived similarity to individual targets by providing participants with cues about shared or unshared attributes. For instance, some participants learned that a target shared their preference for a painting by Klee over one by Kandinsky or that a target shared their guilty verdict in a hypothetical criminal case. Other participants learned that the target preferred the Kandinsky painting or issued an innocent verdict. As expected, these specific cues affected general judgments about similarity. Those who perceived more similarity to a target showed higher levels of projection and lower levels of stereotyping in inferences about an unrelated domain, intuiting the target's competitive motives in a negotiation. In this and other studies, perceived general similarity appeared to moderate the use of inferential strategies in mindreading (Ames, 2004a).

Similar effects emerge for inferences about *group* mental states (Ames, 2004b). In one study, urban university student participants were asked to write about their general similarities to, or differences from, suburban adolescents of the same sex. Those writing about similarities saw themselves as generally more similar to the target group than those writing about differences. In a subsequent task, participants indicated how much they and the target group would like a series of movies based on hypothetical plot summaries. Those who focused on similarities to the group engaged in higher levels of projection for the movie preferences, assuming the group members' preferences would more closely parallel their own tastes. Those who had previously focused on differences engaged in higher levels of stereotyping, assuming adolescent females would show greater preference for movies consistent with a widespread female stereotype (stressing personal growth, dialogue, romance, and sad scenes) while males would show greater preference for movies consistent with a widespread male stereotype (stressing violence, action, nudity, and slapstick comedy).

In sum, research on individual and group-level mindreading has shown a negative relationship between projection and stereotyping. While it is not necessarily or always the case that projection and stereotyping function as alternative strategies that displace each other, such a tension may often emerge. Subjective perceptions of general similarity, whether accurate or not, appear to at least partly guide when these tools are used in mindreading.

EVIDENCE-BASED AND EXTRATARGET STRATEGIES IN INTERACTION

In the preceding sections, I've suggested contingencies within behavioral evidence-based strategies (acts versus emotional displays) and within extratarget strategies (projection versus stereotyping), but how are these *classes* of strategies employed and traded off against one another in mindreading? Prior work in related domains suggests that a range of factors may be involved (see Beike & Sherman, 1994). Drawing on this work, I'll offer a rather general third contingency:

> *Cumulative behavioral evidence supersedes extratarget strategies: projection and stereotyping will drive mindreading when behavioral evidence is ambiguous, but as apparent evidence accumulates, inductive judgments will dominate.*

Several existing models suggest that extratarget strategies (projection, stereotyping) may be a default starting point for much of social inference. Notably, Fiske and Neuberg's (1990) and Brewer's (1988) influential models of stereotyping suggest that social category-based reasoning is a default or initial stage of judgment. According to these and similar accounts, individuating information about a target and his or her behavior are taken into account only when conditions (time, cognitive resources, motivation, and so forth) permit. Like much of stereotyping research, these models revolve around general trait impressions of targets rather than mindreading per se. Yet, given that impression formation is often a matter of mindreading (e.g., to judge if someone is a helpful person, I may first intuit whether he or she has helpful intentions), it seems reasonable to think these same stereotype-by-default effects might extend to mental state inferences.

Other accounts suggest a projection-by-default effect. Recently, Krueger (2003) suggested that "When the responses of others are not known, people project their own as a first bet" (p. 589). Similarly, Epley, Keysar, Van Boven, and Gilovich (2004) have offered an anchoring-and-adjustment model of perspective taking, suggesting that social projection serves as an initial estimate for others' mental states with subsequent and often inadequate adjustments made from this point (see also Barr & Keysar, Chapter 17, and Van Boven & Loewenstein, Chapter 18, in this volume). Their research has shown that time pressure may reduce these adjustments while accuracy incentives may increase them.

Thus, extratarget starting points—whether based on a stereotype or on the self—may be common for mindreaders. There are also reasons to expect that when mindreaders carry such prior assumptions into interactions with targets, these assumptions can become self-fulfilling. Snyder and Swann's (1978) classic work on confirmation processes in impression formation suggests that perceivers may disproportionately elicit and attend to confirming evidence from targets. Likewise, Kelley and Stahelski's (1970) research on social dilemmas describes how competitive players can "assimilate" their cooperative partners: in mistakenly assuming their dove-like partners have hawk-like intentions, competitive players can behave in such a way as to give their cooperative partners little choice but to compete, thereby seemingly confirming the competitive player's initial misguided assumption.

Despite the power of initial expectations and the presence of such self-fulfilling prophecies, initial impressions do change, and the effects of initial mindreading are not entirely rigid. One reason may be a decrease in the potency of stereotypes over the course of an interaction. Recent evidence from Kunda and colleagues (e.g., Kunda, Davies, Adams, & Spencer, 2002) suggests that stereotype activation may dissipate over relatively short periods of time. In one study, after several seconds of observing an interview with a black target person, nonblack participants showed implicit cognitive activation of widely held stereotypes of blacks; after 12 minutes of observation, however, no such activation was evident (Kunda et al., 2002).

Thus, activation of stereotypes may naturally fade, leading to a waning effect on mindreading. Beside such dissipation effects, compelling behavioral evidence that runs counter to stereotypes may override stereotype-driven mindreading. For instance, Krueger and Rothbart (1988) showed that at low levels of behavioral evidence strong stereotypes clearly shaped impressions (e.g., construction workers were seen as more aggressive than housewives). However, when evidence was strong (e.g., consistent aggressive behavior), impressions shifted across the board, and stereotype effects became nonsignificant. Elsewhere, Weisz and Jones (1993) argued that perceivers show a readiness to relinquish category-based expectancies in the face of contradicting behavioral information, in part because of perceivers' willingness to subtype individuals (i.e., continuing to believe the stereotype is still generally true, just not for the focal individual). Flynn, Chatman, and Spataro (2001) found that demographically different work partners were less likely to be negatively stereotyped when they had higher levels of extraversion; the researchers suggested that this effect emerged because extraverted targets provided more stereotype-disconfirming evidence. As with the stereotyping research discussed earlier, these findings have focused on general trait impressions, but it seems likely such effects would extend to mental state inferences.

Stereotypes may thus give way, in part or whole, to accumulating evidence—but what of projection? The question appears to have received only limited attention from social psychologists. In one set of studies, Krueger and Clement (1994) found that, when estimating others' behavior, participants anchored on their own behavior and gradually, though suboptimally, adjusted their estimates as they were provided with information about an increasing number of other cases. Some developmental evidence addresses the emerging ability of children to shift away from projection in light of behavioral evidence. Repacholi and

Gopnik (1997) matched children with an adult experimenter who, contrary to the near universal preference of children, showed disgust when presented with Goldfish snack crackers and displayed delight when presented with broccoli flowerets. When 14-month-olds played a give-and-take game in which the experimenter asked "Can you give me some?" nearly all seemingly projected their own preferences and offered up crackers. By 18 months of age, however, nearly all children deferred to the behavioral evidence and, against their own tastes, offered the broccoli.

It's worth noting, at least in passing, other factors that may affect the balance of power between behavioral-evidence and extratarget mindreading strategies (see Beike & Sherman, 1994, for a review). Effortful examination of behavioral evidence may be stimulated by interaction goals and self-relevancy (e.g., Brewer, 1988) as well as accountability for one's judgments (e.g., Tetlock, 1983). Such consideration of behavioral evidence may be inhibited by cognitive load (e.g., Gilbert & Hixon, 1991), time constraints (e.g., Epley et al., 2004), and social power (e.g., Fiske, 1993).

NEGATIVE AND POSITIVE INFORMATION

What do perceivers most want to know when they read others' minds? Mindreaders are not agnostic onlookers, equally interested in all manner of mental states. On the contrary, perceivers are highly motivated to figure out the interactional and relational motives of the person they're dealing with. For instance, I am vastly more interested in knowing if someone is discreetly trying to take advantage of me as opposed to, say, whether they like my favorite politician or prefer my hair parted on the right rather than the left. Perceivers are not only keen to know someone's social motives in general, but they are also especially vigilant about whether the target has *harmful* motives toward *them*.

This notion is reflected in a variety of work which suggests that, in general, "bad is stronger than good" in psychological life (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001) and that negative information, especially negative *moral* information, is more attention-grabbing (Pratto & John, 1991) and is weighed more heavily in impressions (e.g., Reeder & Coovert, 1986; Ybarra, 2001). This sentiment also seems to be consistent with work on cheater detection and the Machiavellian intelligence hypothesis (e.g., Whiten, 1997), which variously suggests that the evolution of theory of mind may have been partly spurred by the need to defend against exploitation by duplicitous conspecifics. Thus, another plausible contingency is:

> *Negative social intention information weighs heavily in mindreading: within a mindreading strategy, cues signaling negative social intentions may dominate neutral or positive cues; between mindreading strategies, those strategies that signal negative social intentions may dominate.*

The within-strategy effect may be most meaningful for stereotypes and behaviors. For instance, one could imagine that, upon meeting an ex-convict librarian, a perceiver's mindreading may be more swayed by the ex-convict stereotype than the librarian one. As for behaviors, previous work (e.g., Reeder & Coovert, 1986) suggests that negative behaviors worsen positive initial impressions more than positive behaviors improve negative initial impressions.

There does not appear to be much work on mindreading that tests for negative evaluation effects *between* inferential strategies. Nonetheless, it is tempting to speculate that a negative stereotype might be particularly resistant to positive behavioral evidence. Conversely, negative behaviors might quickly swamp a positive stereotype. One mechanism underlying such an effect could be attention. For instance, Fox and colleagues have shown that perceivers are faster to detect, look longer at, and are more reluctant to disengage from, angry faces (Fox, Russo, & Dutton, 2002). It's an empirical question, but one can readily imagine that a perceiver confronted by a snarling nun might instinctively defer to behavioral evidence, whereas a perceiver confronted by a cheerful soldier of fortune might nonetheless cling to stereotypes.

This contingency concerning negative social information is admittedly speculative and rather general. Whether or not it will survive additional testing and elaboration remains to be seen, but behind this contingency lies an important pair of questions deserving further thought, namely, "Which kinds of mental states are mindreaders most interested in intuiting" and "How do these priorities shape the mindreader's use of cues and inferential strategies?"

CONCLUSION

Scholarly accounts of mindreading may have become so numerous and compelling that, in some ways, the question is no longer "how do people read minds?" but rather "how *don't* they?" in any given instance. Researchers have mapped out much of the perceiver's toolkit, but have only begun to account for which tool is used when.

Several important strategies seem to account for much of everyday mindreading. Some of these revolve around direct evidence from the target (scrutinizing behaviors in context; reading emotional displays) while others rely on extratarget information (applying stereotypes; projecting one's own beliefs, desires, and feelings). It may be that these strategies are often used in combination, as when we rely on a stereotype to interpret an emotional display. Yet, I suggest perceivers frequently shift between these strategies, which begs questions about when and why such shifts occurs.

In this chapter, I offered four contingencies about when various inferential tools might be used. I suggested that affect qualifies behavior in the short term, that perceived similarity governs projection and stereotyping, that cumulative behavioral evidence supersedes extratarget strategies, and that negative social intention information weighs heavily in mindreading. These contingencies have varying degrees of empirical support and would benefit from additional research and thinking. Additional or alternative contingencies are certainly possible. Regardless of the content of a contingency, though, important associated questions emerge, including:

- *What's the developmental course of the contingency?* Perhaps 5-year-olds shift strategies differently from 50-year-olds.
- *What are the consequences of the contingency?* The contingency may entail benefits in terms of speed or accuracy as well as costs in terms of effort and distortion.
- *What cognitive mechanisms could instantiate the contingency?* Some contingencies might themselves be inferential rules that are implicitly represented, while others imply cognitive processes, such as parallel-constraint satisfaction and coherence mechanisms (e.g., Kunda & Thagard, 1996; Read & Miller, 1998) or anchoring-and-adjustment (e.g., Epley et al., 2004).

Though the present space does not permit it, each of these issues could be fruitfully examined for the contingencies offered here. These same questions apply to other contingencies—and I hope more such proposals will soon be offered. The story of everyday mindreading will not be complete until we can answer the question of "Which tools are used when?" This chapter sketches some seemingly promising possibilities; I look forward to seeing more responses emerge.

NOTE

1. Several other meaningful strategies are not portrayed in Figure 10.1. One is reasoning by *analogy*. As various scholars have argued (e.g., Andersen,

Glassman, Chen, & Cole, 1995), we may use familiar others as templates for new acquaintances. Such effects seem likely to extend to mindreading: a blind date may remind us of a past romantic partner, and so we may assume he or she has the same tastes, passions, and pathologies as our "ex." More generally, perceivers may consult *generic prototypes*, representations of what the prototypical person would think and feel in a given situation (e.g., Karniol, 2003). Another strategy is the reliance on well-developed *prior impressions* of specific individuals in our subsequent interactions with them (see, e.g., Weisz & Jones's [1993] discussion of target-based expectancies). Such person-specific theories may guide much of our daily mindreading with close others.

REFERENCES

Ames, D. R. (2004a). Inside the mind-reader's toolkit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology, 87*, 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 and Social Psychology, 87*, 573–585.

Ames, D. R., Johar, G. V., & Kammrath, L. K. (2004). *The impact of affective displays on impression formation: I know what you're like when I see how you feel*. Unpublished manuscript.

Andersen, S. M.. Glassman, N. S., Chen, S., & Cole, S. W. (1995). Transference in social perception: The role of chronic accessibility in significant-other representations. *Journal of Personality and Social Psychology, 69*, 41–57.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323–370.

Beike, D. R., & Sherman, S. J. (1994). Social inference: Inductions, deductions, and analogies. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 209–285). Hillsdale, NJ: Lawrence Erlbaum.

Brewer, M. B. (1988). A dual process model of impression formation. In R. Wyer & T. Srull (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Erlbaum.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*, 327–339.

Fiske, S. T. (1993). Controlling other people: The impact of power on stereotyping. *American Psychologist, 48*, 621–628.

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. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). New York: Academic Press.

Flynn, F. J., Chatman, J. A., & Spataro, S. E. (2001). Getting to know you: The influence of personality on the impression formation and performance of demographically different people in organizations. *Administrative Science Quarterly*, 46, 414–442.

Fox, E., Russo, R., & Dutton, K. (2002). Attentional bias for threat: Evidence for delayed disengagement from emotional faces. *Cognition and Emotion, 16*, 355–379.

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.

Goffman, E. (1959). *The presentation of self in everyday life*. New York: Anchor Books.

Heider, F. (1958). *The psychology of interpersonal relations*. Hillsdale, NJ: Erlbaum.

Hugenberg, K., & Bodenhausen, G. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14*, 640–643.

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). New York: Academic Press.

Kammrath, L., Mendoza-Denton, R., & Mischel, W. (in press). What's in a trait? Mental states and If . . . then . . . profiles in folk theories of traits. *Journal of Personality and Social Psychology.*

Karniol, R. (2003). Egocentrism versus protocentrism: The status of self in social prediction. *Psychological Review, 110*, 564–580.

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation, 14*, 192–241.

Kelley, H. H., & Stahelski, A. J. (1970). Errors in perception of intentions in a mixed-motive game. *Journal of Experimental Social Psychology, 6*, 379–400.

Krueger, J. (1998). On the perception of social consensus. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 163–240). San Diego, CA: Academic Press.

Krueger, J. (2003). Return of the ego—self-referent information as a filter for social prediction: Comment on Karniol (2003). *Psychological Review, 110*, 585–590.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality & Social Psychology, 67*, 596–610.

Krueger, K., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology, 55*, 187–195.

Kunda, Z., & Davies, P. G., Adams, B. D., & Spencer, S. J. (2002). The dynamic time course of stereotype activation: Activation, dissipation, and resurrection. *Journal of Personality and Social Psychology, 82*, 283–299.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*, 284–308.

Malle, B. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Meltzoff, A. N., & Brooks, R. (2001). "Like me" as a building block for understanding other minds: Bodily acts, attention, and intention. In B. Malle, L.

Moses, & D. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 171–191). Cambridge, MA: MIT Press.

Newtson, D. (1974). Dispositional inference from effects of actions: Effects chosen and effects foregone. *Journal of Experimental Social Psychology, 10*, 489–496.

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*, 380–391.

Pollick, F. E., Paterson, H. M., Bruderlin, A., & Sanford, A. J. (2001). Perceiving affect from arm movement. *Cognition, 82*, B51–B61.

Read, S. J., & Miller, L. C. (1998). On the dynamic construction of meaning: An interactive activation and competition model of social perception. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and behavior* (pp. 27–68). Mahwah, NJ: Erlbaum.

Reeder, G. D., Kumar, S., Hesson-McInnis, M., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology, 83*, 789–803.

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition, 4*, 1–17.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14– and 18–month-olds. *Developmental Psychology, 33*, 12–21.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39*, 590–598.

Snyder, M., & Swann, W. B. (1978). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology, 14*, 148–162.

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45*, 74–83.

Van Boven, L., & Loewenstein, G. (in press). Varieties of social projection. To appear in M. Alicke, D. Dunning, & J. Krueger (Eds.), *The self in social perception*. Hove, UK: Psychology Press.

Weisz, C., & Jones, E. E. (1993). Expectancy disconfirmation and dispositional inference: Latent strength of target-based and category-based expectancies. *Personality and Social Psychology Bulletin, 19*, 563–573.

Whiten, A. (1997). The Machiavellian mindreader. In A. Whiten & R. W. Byrne (Eds.), *Machiavellian intelligence: Vol. 2. Extensions and evaluations* (pp. 144–173). Cambridge, UK: Cambridge University Press.

Ybarra, O. (2001). When first impressions don't last: The role of isolation and adaptation processes in the revision of evaluative impressions. *Social Cognition, 19*, 491–520.

11

Mental Simulation

Royal Road to Other Minds?

JOSEF PERNER
ANTON KÜHBERGER

Simulation theory deals with how we interpret and understand other minds. Historically, simulation theory has ancestors in German psychology (Lipps, 1907) and philosophy (Scheler, 1973; see Schäfer, 1980, for Husserl) and the English tradition of historical understanding (Collingwood, 1946/1993; see also Blackburn, 1992). In 1986, Jane Heal, under the name of "replication," and Bob Gordon, under the name of "simulation," introduced these ideas into the recent philosophical discussion as an alternative to the dominant "theory" view of understanding other minds (e.g., Churchland, 1984), in which mental terms are theoretical constructs in a theory of behavior (Sellars, 1956). The original controversy between the two camps (e.g., Goldman, 1989; Stich & Nichols, 1992), based on all-or-none positions, mustered empirical evidence from social and developmental psychology (Davies & Stone, 1995a, 1995b). These theoretical positions, however, turned out to be too flexible to be constrained by the available evidence. Moreover, the philosophical discussion resulted in the compromise that we use both kinds of processes (e.g., Heal, 1995; Stich & Nichols, 1995), which makes it even more urgent to find empirical constraints in the two theories. We suspect that this is particularly difficult for the extremely flexible theory theory and has more of a chance in the case of simulation theory. Our aim here is to formulate explicit principles from which we can deduce empirically testable implications concerning the accurate human prediction of mental states or behavior by simulation.

SIMULATION VERSUS THEORY

In general, the term "simulation" is applied when trying to understand or predict the behavior of some system for which there is no theory of how it functions. Instead, knowledge about the system is gained by exploring the functioning of an analogue system that has the same causal structure in relevant aspects. That is, the system to be understood is simulated by the analogue system that is manipulable. For instance, before building a real bridge the engineers can construct a small model bridge in such a way that appropriately scaled down water pressures can be used to see whether the actual construction will withstand the actual pressures. This simulation of the effects of water pressures on the model bridge is necessary if no theory can predict them accurately. However, this does not mean that we can do without any theory; for the model will produce analogue effects only if the elements are set up the right way. For instance, the strength of the flow must be scaled down in proportion to the scale of the model, and many other aspects must be set on the basis of theory.

Model bridges can be used for simulating real bridges, and our own mind can be used to simulate the functioning of other minds to the degree to which minds obey the same laws. For instance, if one wants to predict how someone will react to encountering a lion in the wild, one can fly to Africa and check out the effects of such an encounter on one's own mind and behavior. This type of simulation, which is comparable to the simulation of physical processes like water pressures on bridges, has been termed "actual situation simulation" by Stich and Nichols (1997). It is obviously not of breathtaking advantage, since identical situations to run the simulation are needed. *Mental simulation* of the mind, however, adds a unique feature. In order to simulate another person's mental and behavioral reactions to a situation one does not need put oneself into the same situation, but one can do it in imagination. Instead of flying to Africa one can simply imagine oneself encountering a lion in the wild and then experience (in a rather scaled-down fashion) the panic-eliciting quality of this situation. Stich and Nichols call this type "pretense-driven offline simulation." Of course, mental simulation must be theoretically informed when being set up; that is, one needs to know the relevant features to be represented in imagination: a ferocious and hungry lion with nobody around to protect me, and not a sleeping lion watched from an armored car.

Simulation of physical processes is relatively easy to distinguish from a theory about physical processes. Simulation is itself a physical process, while the theory is a mental process. The same distinction becomes tricky for mental simulation, because both theory and simulation are mental processes. This raised the threat that the distinction between mental simulation and theory would collapse (Davies & Stone, 1998; Heal, 1994). This threat can be avoided if we are allowed perusal of the distinction between attitude and representational content inherent in the philosophical characterization of mental states as propositional attitudes (Brentano, 1874/1955; Russell, 1919). In this standard view, a mental state—for instance, the knowledge that one owns a particular object—is analyzed as having the attitude of KNOWLEDGE toward the proposition "I own this object." We use this distinction in Table 11.1 to schematically characterize the sequence of mental states one might run through in an endowment experiment (Thaler, 1980).

Loewenstein and Adler (1995), in an experiment on the endowment effect, gave a group of students a mug with their university's logo emblazoned on it. Another group of students without a mug (no-endowment condition) were asked how much at most they were willing to pay for one. Their average offer was $4.05. The students who had just been given a mug (endowment condition) were asked for how much minimally they were willing to sell it. Their average selling price was considerably higher: $5.96. This buyer–seller discrepancy constitutes the endowment effect.

We use a version of endowment and no-endowment conditions of this experiment to highlight the theoretically important distinction between real events, including the possession of mental states (attitudes

TABLE 11.1. Sequence of Mental States and Situational Changes in an *Actual* Endowment Experiment

No endowment (buyer)	Endowment (seller)
	(1) PERCEIVE [being given mug]
	(2) KNOW [mug is mine]
	(3) FEEL HAPPY [to own mug]
(1) PERCEIVE [asked to buy mug]	(4) PERCEIVE [asked to sell mug]
(2) ESTIMATE [mug is worth $4]	(5) ESTIMATE [my mug is worth $6]
(3) ANSWER[a]: ["I pay $4"]	(6) ANSWER[b]: ["I sell for $6"]

[a] This is the answer to the question "For how much would one buy the mug?"

[b] This is the answer to the question "For how much would one sell the mug?"

and their representations)[1] and what is merely represented (representational content, proposition) by capitalizing the real events (including attitudes and representations) and by putting representational contents between brackets.

The basis for the endowment effect is that being endowed with an object (e.g., a mug) increases its subjective value, which results in a higher selling price than the original offer to buy it. Table 11.2 shows the mental states one has to entertain if one has to predict by using a "theory" how participants in this experiment will answer the two questions about buying and selling prices. The important aspect that makes it a "theory" is the fact that there is knowledge of how people value something and how their valuation of an object increases when being endowed with that object.

Table 11.3 shows how these predictions are made by using simulation. The hallmark of simulation is the absence of any knowledge about how other people tend to act and value objects before or after being endowed with them. Proceeding by simulation, one imagines being in the experiment oneself and observes one's own answer tendencies triggered by this imagination. These triggered answers one ascribes to the participants as their answers.

Comparison of the sequence of mental states in Tables 11.1 and 11.3 highlights one of the critical assumptions of simulation theory: that one undergoes similar mental states as the simulated person, except "offline." Obviously one is not in the same situation as the participants in the experiment, so one cannot *perceive* the situation but only *imagine* that situation—though in some cases one can engage in pretty much the

TABLE 11.2. *Predicting* Participants' Answers in Endowment Experiment by *Theory Use*

No endowment (buyer)	Endowment (seller)
(*Entry*) TOLD ABOUT PARTICIPANTS IN THIS EXPERIMENT AND ASKED HOW THEY WILL RESPOND TO BUYING QUESTION	(*Entry*) TOLD ABOUT PARTICIPANTS IN THIS EXPERIMENT AND ASKED HOW THEY WILL RESPOND TO SELLING QUESTION
(1) KNOW [When someone is asked to buy a mug being worth $4, she (or he) will *offer* $4. The answer will be "She'll buy for $4."]	(1) KNOW [When someone gets possession of something—and *knows* that it is now hers—then she will *feel happy* and come to *value* the object by about one-half more than before. So, a mug worth $4 might be *valued* at $6. The answer will be "She'll sell for $6."]
(*Exit*) ANSWER QUESTION.	(*Exit*) ANSWER QUESTION.

TABLE 11.3. ***Predicting*** **Answers in Endowment Experiment by** ***Mentally Simulating*** **Participants**

No endowment (buyer)	Endowment (seller)
(*Entry*) TOLD ABOUT PARTICIPANTS IN THIS EXPERIMENT AND ASKED HOW THEY WILL RESPOND TO BUYING QUESTION	(*Entry*) TOLD ABOUT PARTICIPANTS IN THIS EXPERIMENT AND ASKED HOW THEY WILL RESPOND TO SELLING QUESTION
	(1) IMAGINE [being given mug]
	(2) ASSUME [mug is mine]
	(3) FEEL SORT-OF-HAPPY [to own mug]
(1) IMAGINE [asked to buy mug]	(4) IMAGINE [asked to sell mug]
(2) ESTIMATE [mug is worth $4]	(5) ESTIMATE [my mug is worth $6]
(3) ANSWER TENDENCY: ["I pay $4"]	(6) ANSWER TENDENCY: ["I sell for $6"]
(*Exit*) ASCRIBE OWN ANSWER TENDENCY TO PARTICIPANTS	(*Exit*) ASCRIBE OWN ANSWER TENDENCY TO PARTICIPANTS

same mental states as one would if one were an actual participant, for example, by *estimating* the subjective value of the mug. However, one cannot *know* that one owns the mug, since one knows that one is just imagining; hence, one can at best *assume* that one owns the mug.

Now, why should the simulated processes of "perceiving (x) → knowing (y) → being happy about (z)" be governed by the same causal processes as "imagining (x) → assuming (y) → feeling sort-of-happy about (z)"? That this should be so is a big assumption to make. Fortunately there is increasing evidence that processes of visual perception do indeed use the same neural architecture as processes of visual imagery (e.g., Gallese & Goldman, 1998; Gallese, Keysers, & Rizzolatti, 2004; Kosslyn & Thompson, 2003). So, one could consider the change from actually perceiving something to merely imagining it, or from really knowing that something is true to merely assuming it is true (pretend believing) as corresponding to the scaling down from actual bridges to model bridges in simulations of physical processes. This view is not totally unrealistic, as suggested by common experience. For instance, merely imagining how the steps of a dubious-looking character are closing in on us in a dark alley elicits very similar feelings and leads to the same quickening of our own pace as when it actually happens.

CONSTRAINING SIMULATION THEORY

We are primarily concerned with the ability to predict how different situations affect people differently. We need to distinguish three elements:

the situation the person is in, how the situation affects the person (the causal link), and the consequences of the causal process on the person. Each of these three elements can be composed of physical and mental elements. As a crude example, let us take the situation in which a person gets attacked by a robber with a metal club. The relevant physical aspects of the situation consist of the presence of the robber and his actions. Mental aspects of the situation consist of the moral depravity of his actions, his intentions, and so on. The consequences for the victim can also be physical and mental. The immanent attack will probably trigger fear (mental) and sweating (physical), and after the club hits the skull there may be damage to the brain (physical) with altered states of consciousness in its wake (mental). Of central importance for our attempts to constrain simulation theory empirically is the fact that the causation itself—the way in which this situation causes these consequences—can also be physical or mental (physical vs. mental causation). The immanent attack triggers fear and sweating mentally, because it only works if the victim perceives (has a mental state about) what is going to happen to him or her. If the robber approached while the victim was already unconscious (for some other reason) he or she would not suffer these consequences. In contrast, the blow to the head is purely physical causation, insofar as it would have the same effects regardless of whether the victim was conscious or unconscious at the time of the attack.

The cornerstone of simulating the mind of another person is that one imagines oneself being in the other person's situation, which hopefully will trigger off-line processes that are similar to those of the other. Now, one can imagine physical and mental aspects of the situation. This can have physical or mental consequences, but—and this is the critical constraint—these consequences must be mentally caused for simulation to work. Physically caused consequences cannot be simulated. That is, I can only simulate the consequences on the robber's victim that are mediated by the victim perceiving what is happening (mental causation). Living through this experience in imagination can give me fear and make me sweat. I cannot simulate the physically caused consequences of the blow on the head. My imagining receiving a blow to my head will not trigger, not even offline, a minor version of the victim's brain damage and its ensuing neuropsychological syndromes (mental effects). This constraint on what can and cannot be simulated we formulate as a first principle:

> *Simulation Principle #1: "Mental Causation"*
>
> *The consequences of situations can be mentally simulated only if the consequences are due to the situation being represented by a mental state. Purely physical (nonmental) influences cannot be mentally simulated.*

As an instructive example let us consider the question of whether the mental and behavioral consequences of alcohol consumption can be correctly predicted on the basis of simulation (Davies & Stone, 1998; Perner & Kühberger, 2003; this example assumes that the predictor has never consumed alcohol and fails to have any expectations about its possible effects). Is it possible to accurately predict how a person will behave after we saw him (for example) drinking five bottles of strong beer by using simulation? We first need to analyze the mental states of the drinker. The drinker knows, of course, what he was doing because he perceived himself drinking the beer. We can simulate this perception of the situation in scaled-down form by imagining ourselves drinking the five beers. However—and this is the critical point here—from this imagination no relevant scaled-down symptoms of drunkenness follow, because the actual symptoms of drunkenness of the drinker are not a consequence of his perceptions of his beer drinking, but rather are a physical effect of alcohol ingestion. This can be seen from the following thought experiment (see Table 11.4). In a 2 × 2 design we manipulate as factors "reality" (i.e., whether the drinker consumes five regular beers or five alcohol-free beers) and "belief" (i.e., whether the drinker believes he is drinking five regular beers or five alcohol-free beers).

Unless our intuition is completely mistaken, the person will behave drunkenly after having consumed five real beers, and will show no signs of drunkenness after consumption of five alcohol-free beers—*regardless* of what he thinks he was drinking (given that the drinker fails to have any expectations about the effects of his drinking). That is, the behavioral consequences are solely due to the physical cause of alcohol on the mind and are not due to the mental cause of believing that one was consuming alcohol. Consequently, according to our "mental causation" principle, the behavioral consequences of alcohol consumption cannot be accurately predicted by using simulation.

We have to distinguish between the behavioral consequences of alcohol consumption and the behavioral consequences of being and/or feeling drunk (see Figure 11.1). Alcohol consumption comes first, resulting in being and, usually, feeling drunk, which in turn leads to the typical

TABLE 11.4. Expected Behavioral Consequences in Beer Experiment

	Belief	
Reality	5 strong beers	5 alcohol-free beers
5 strong beers	*Drunken*	*Drunken*
5 alcohol-free beers	Sober	Sober

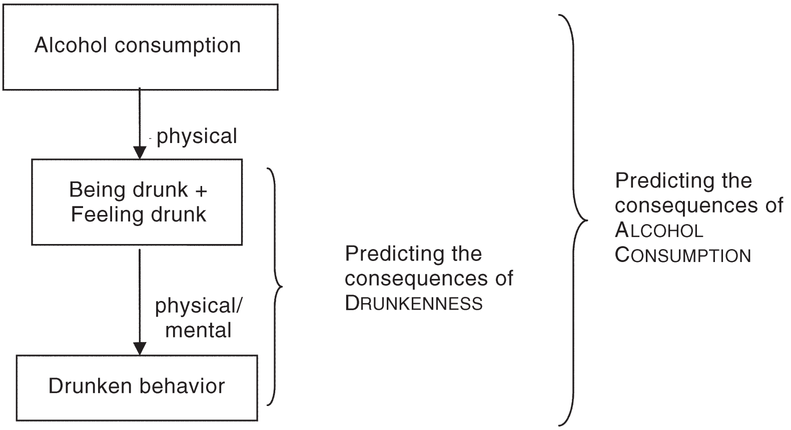
Alcohol consumption
physical
Being drunk + Feeling drunk
physical/ mental
Drunken behavior
Predicting the consequences of DRUNKENNESS
Predicting the consequences of ALCOHOL CONSUMPTION

FIGURE 11.1. Distinguishing effects of drunkenness from those of alcohol consumption.

drunken behaviors. What we claim is (shown at the right side of Figure 11.1) that simulation cannot link alcohol consumption to drunken behavior. In other words, a person not familiar with the experience of being drunk will not be able to predict the typical behaviors of a drunken person by using simulation.

This conclusion appears, however, to contradict some people's intuition that simulation of drunken behavior is possible. This intuition rests on the polysemy of "simulation." Here, it is used to mean that we can pretend to be drunk. But that would be pretend behavior based on knowledge (theory) of how drunk people behave. No simulation in our sense is involved in this. What we need to distinguish simulation from theory is that the imagination of drinking the beer leads one to an experience of wanting to act in a certain way (which we recognize as that of a drunkard's) without any knowledge that drinking beer leads to such behavior. Figure 11.1 also shows a reading of "simulating drunken behavior" (middle parenthesis) that does involve simulation in the relevant sense. However, it is not simulation of the consequences of alcohol consumption but simulation of the consequences of drunkenness.

The effect of alcohol consumption as a physical effect on the mind stands in instructive contrast to the endowment effect, which clearly is a mental effect. This highly replicable effect is particularly interesting, because data by Loewenstein and Adler (1995) suggested that it cannot be predicted by ordinary people in an experiment. However, according to our principle of mental causation, it should actually be correctly predicted via simulation. To see whether it is a mentally or physically caused effect let us again consider a thought experiment on the endowment effect, in which we add to the two reality conditions two belief conditions: (1) students own the mug but think they do not own it, or (2) they are wrongly told the mug is theirs when it really isn't. Our clear intuition is that what matters is the students' belief and not the actual possession, as Table 11.5 summarizes.

Loewenstein and Adler (1995) tested whether students who were not given a mug could predict how much a student in the endowment condition would ask for selling a mug. Their mean predicted selling price looked practically the same ($4.16) as the amount they were willing to spend on a mug ($4.05). How could simulation account for this prediction failure? Fortunately there is no real need for an explanation because a nonsignificant small difference does not preclude the existence of a difference in the expected direction. Indeed, a more extensive study by Van Boven, Dunning, and Loewenstein (2000) replicated that sellers or buyers were not very accurate in predicting the offers or asking prices by people in the other role, but they did show a difference in the right direction, and studies with lottery tickets instead of mugs showed even more accurate predictions of buying and selling prices in the endowment experiment (Gilhofer, 2002; Knetsch & Sinden, 1984; see Perner & Kühberger, 2003).

Gilhofer (2002) also showed a *duration of endowment effect* using lottery tickets. People wanted even more for their ticket when they had been in possession of it for a full week than when asked to sell it a few minutes after they had been endowed with it (see Strahilevitz & Loewenstein, 1998, for shorter duration effects with mugs). In order to answer the question whether people can accurately predict this duration of endowment effect by using simulation, we need to formulate a second principle.

Simulation Principle #2: "Mental State Features"

Effects of features of mental states can be accurately simulated only if the states evoked in simulation have the same or similar features.

TABLE 11.5. Expected Subjective Valuations in Endowment Thought Experiment

	Belief	
Reality	Endowed with mug	Not given mug
Endowed with mug	*High Value*	Low Value
Not given mug	*High Value*	Low Value

For duration this means that effects of duration can be accurately simulated only if the corresponding state in the simulative process has a similar duration. On the assumption that the duration of endowment effect rests on the duration of actually knowing about the endowment, we can try to answer the question whether this effect could be predicted by using simulation. On the basis of our principle #2 (mental state feature) the answer is clearly "no," if one has but a few minutes for making the prediction. To simulate correctly one would need about a week to carry out the simulation—that is, one would have to assume (as a scaled-down version of actually knowing) for a whole week that one owns the ticket. In fact, Gilhofer (2002) gave people only a few minutes to predict what someone owning the ticket for a week would ask as a selling price. Predictions did not reflect the duration of endowment effect but were the same (or even slightly less) than predictions for the immediate endowment effect (see Perner & Kühberger, 2003).

So, what we have achieved so far is to derive two empirical implications of simulation theory:

1. Physical effects of situations (e.g., effects of alcohol consumption) on the mind and on behavior cannot be accurately predicted by using simulation, while mental effects of perceiving situations can—in principle.
2. Effects that are due to properties of mental states, like their duration, cannot be predicted on the basis of simulation unless the simulative process has the same or a similar feature (e.g., a similar duration).

Moreover, available data tend to show that accurate predictions are possible in those cases where accurate prediction should be possible on the basis of simulation (e.g., endowment effect). We also found that in those cases, where simulation should not yield accurate predictions, people actually failed to predict correctly (e.g., duration of endowment).

This looks like impressive evidence in favor of simulation theory. However, the underlying claim that people base all their predictions on simulation may be in the long run very short-sighted. There is no reason why people could not also use theory on some occasions; and in all likelihood they do. For instance, if we find that people can make accurate predictions under conditions when accurate simulation should not be possible, are we to conclude that simulation theory is wrong or that people happen to have a correct theory? This question can be tackled by a method developed by Perner, Gschaider, Kühberger, and Schrofner (1999).

TESTING FOR USE OF SIMULATION VERSUS THEORY

The basic idea behind the method developed by Perner and colleagues (1999) centers on two features.

1. It puts the focus on contrasting pairs of situations (e.g., endowment vs. no-endowment condition) in order to single out a particular aspect (e.g., the endowment), so that we can decide whether the effects of this particular aspect are captured by simulation or theory.
2. It contrasts prediction accuracy under two different ways of presenting the relevant situations to the predictors.

In the *independent prediction* condition each participant is asked for predictions for only one of the two experimental conditions (e.g., one group of people has to make predictions for the endowment condition, another group for the no-endowment condition). In the *juxtaposed prediction* condition each participant is asked to make a prediction for both experimental conditions presented side by side (e.g., each participant is asked to predict what a person in the endowment condition would do and what a person in the no-endowment condition would do).

According to simulation theory, the accuracy of a prediction depends on how well the imagined situation triggers the same processes in the imagining person as those that occur in the person who experiences the situation in real life. On face value, it seems plausible that imagining a single situation would serve this purpose optimally. However, when asked to simulate two different situations at the same time, this might (but need not necessarily) lead to interference. In any case, it is hard to see how executing two predictions at a time (one after the other or really simultaneously) could lead to an improvement of the prediction when using simulation. On the basis of these considerations we suggest:

> *If predictions for two juxtaposed situations are more accurate than independent predictions for each single situation, then simulation theory has no ready explanation at hand.*

In contrast, theory-theory has a ready explanation for why predictions for juxtaposed situations would be more accurate than predictions for each situation in isolation: the juxtaposition singles out the relevant factor that produces the target effect (i.e., it produces a focusing effect; see Schkade & Kahneman, 1998). Hence, the theory can be applied accurately without being distracted by other potentially relevant factors applicable to each situation in isolation. With these considerations we can refine our empirical implications of simulation theory (ST) as follows:

> *If*, according to ST, accurate predictions are possible,
> *Then* a positive difference should exist in independent predictions. (Otherwise a plausible argument is needed that a wrong theory is interfering.)
>
> *If*, according to ST, accurate predictions are not possible,
> *Then either* independent predictions should show a zero difference,
> *Or*,
> *If* independent predictions do show a positive or negative difference,
> *Then* juxtaposed predictions must show at least an equally large difference (which makes the positive difference attributable to the use of a theory).

We have used the two prediction conditions on a variety of tasks, and the resulting picture is very much in tune with simulation theory. Consider first tasks, where simulation should be able to yield accurate predictions. For instance, participants were able to predict the influence of rating scales of different length on estimates of vaguely known facts (e.g., the top speed of a good race horse). Although juxtaposition yielded an even larger difference than independent predictions (Perner et al., 1999: indicating the use of a theory), independent predictions also showed a significant positive difference, which is compatible with the use of simulation in addition to theory use. Similarly, Perner and Kühberger (2002) found that independent predictions of choices in a gambling version of the famous framing task (Tversky & Kahneman, 1981: negatively framed options lead to more risky choices than positively framed options) differed in the direction of the actual effect, and so did predictions for a payoff effect (riskier choices for low payoffs than for high payoffs). As reported above, independent predictions for the endowment effect differed in the right direction (Gilhofer, 2002; Knetsch & Sinden, 1984; Van Boven et al., 2000). The same was found for the position bias, which is the finding that people tend to choose the rightmost item in an array of identical items (Nisbett & Wilson, 1977) when the critical conditions are described to the predicting participants (Kühberger, Kogler, Hug, & Mösl, 2004). Finally, it was claimed that the effect described by Langer (1975)—that people value a lottery ticket more if they can choose from among three tickets than if they are given just the one ticket—could not be predicted by ordinary people (Nichols, Stich, Leslie, & Klein, 1996). Since there is no reason why this should not be correctly predicted by simulation and why simulation could not be used in this case, this would pose a problem for simulation theory. Fortunately for simulation theory, we could not find any proper replication of Langer's original report and failed ourselves to replicate the effect in four experiments (Kühberger, Perner, Schulte, & Leingruber, 1995). Where there is no reliable target effect we cannot expect correct predictions even if simulation theory implies that an effect—if there were one—should be predictable.

There is one effect, the duration of endowment effect (Gilhofer, 2002; Strahilewitz & Loewenstein, 1998), that people should be incapable of predicting using simulation, following our principle #2. And in line with simulation theory we found that predictions showed a zero difference (or even slightly inverse difference from the target effect; Perner & Kühberger, 2003). Several other duration effects in the literature confirm this picture (see, e.g., the findings on duration neglect; Fredrickson & Kahneman, 1993; Kahneman, Fredrickson, Schreiber, & Redelmaier, 1993).

The most interesting test of simulation theory as implemented by our two principles is the case of an effect that simulation theory says cannot be accurately predicted by simulation but that people, nevertheless, are able to accurately predict. In this case simulation theory can be saved only if it turns out that the correct prediction is based on theory. We can test this by investigating whether prediction under juxtaposed presentation of conditions is better than independent predictions for each condition individually (or at least independent predictions must not be better than juxtaposed predictions). Unfortunately we have not yet found a feasible test case apart from our thought experiment with alcohol consumption. Plagued by doubts that we can implement the required conditions in a theoretically satisfactory way and reluctant to present our design to the ethics commission, we have not carried out this experiment yet.

CONCLUSION

Simulation theory and theory-theory posit quite distinct approaches to people's understanding of other minds. In particular, mental simulation provides a unique access to the mind that makes understanding of the mind different from understanding of everything else. The consensus among philosophers, who have proposed this fundamental distinction, is that people use both theory and simulation in their understanding of the mind. We have developed two principles of simulation theory that make the theory amenable to empirical testing in combination with two different prediction conditions that help decide whether a prediction is based on simulation or on theory. The data so far are remarkably consistent with the assumption that simulation is used whenever it can be profitably used (in addition to occasional "theoretical" knowledge). Many of the contributions to this volume bear witness to the fact that an increasing number of researchers from different areas find simulation a useful concept for explaining people's ability to understand other minds.

NOTE

1. According to the representational theory of mind (Field, 1978; Fodor, 1978), propositional attitudes are implemented in our brain by physical representations (neural activities) whose representational content is the proposition. The representations interact in particular ways with the environment, actions, and other representations. These interactions define the attitude. Our notation separates the physical (environment, action, representation, and interactions between these elements) from the representational content (represented proposition).

REFERENCES

Blackburn, S. (1992). Theory, observation and drama. *Mind and Language, 7*, 187–203.

Brentano, F. von (1955). *Psychologie vom empirischen Standpunkt*. Hamburg: Felix Meiner. (Original work published 1874)

Churchland, P. M. (1984). *Matter and consciousness: A contemporary introduction to the philosophy of mind*. Cambridge, MA: MIT Press.

Collingwood, R. G. (1993). *The idea of history.* New York: Oxford University Press. (Original work published 1946)

Davies, M., & Stone, T. (1995a). *Folk Psychology: The theory of mind debate*. Oxford, UK: Blackwell.

Davies, M., & Stone, T. (1995b). *Mental simulation: Evaluations and applications.* Oxford, UK: Blackwell.

Davies, M., & Stone, T. (1998). Folk psychology and mental simulation. In A. O'Hear (Ed.), *Current issues in the philosophy of mind* (pp. 53–82). Cambridge, UK: Cambridge University Press.

Field, H. (1978). Mental representation. *Erkenntnis, 13,* 9–61.

Fodor, J. A. (1978). Propositional attitudes. *The Monist, 61,* 501–523.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluation of affective episodes. *Journal of Personality and Social Psychology, 65*, 44–55.

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

Gallese, V., & Goldman, A. (1998). Mirror neurons and the Simulation theory of mindreading. *Trends in Cognitive Sciences*, *2*, 493–501.

Gilhofer, R. (2002). *Endowment Effekte in realen und hypothetischen Situationen*. Thesis, University of Salzburg.

Goldman, A. I. (1989). Interpretation psychologized. *Mind and Language, 4,* 161–185.

Heal, J. (1994). Simulation vs. theory theory: What is at issue? *Proceedings of the British Academy, 83,* 129–144.

Heal, J. (1995). How to think about thinking. In M. Davies & T. Stone (Eds.), *Mental simulation* (pp. 33–52). Oxford, UK: Blackwell.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmaier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*, 401–405.

Knetsch, J. L., & Sinden, J. A. (1984). Willingness to pay and compensation demanded: Experimental evidence of an unexpected disparity in measures of value. *Quarterly Journal of Economics*, *99*, 507–521.

Kosslyn, S. M., & Thompson, W. L. (2003). When is early visual cortex activated during visual mental imagery? *Psychological Bulletin*, *129*, 723–746.

Kühberger, A., Kogler, C., Hug, A., & Mösl, E. (2004). *The role of the position effect in theory and simulation*. Manuscript submitted for publication.

Kühberger, A., Perner, J., Schulte, M., & Leingruber, R. (1995). Choice or no choice: Is the Langer effect evidence against simulation? *Mind and Language*, *10*, 423–436.

Langer, E. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32,* 311–328.

Lipps, T. (1907). *Psychologische Untersuchungen*. Leipzig: Engelmann.

Loewenstein, G., & Adler, D. (1995). A bias in the prediction of tastes. *The Economic Journal*, *105*, 929–937.

Nichols, S., Stich, S., Leslie, A. M., & Klein, D. (1996). Varieties of off-line simulation. In P. Carruthers (Ed.), *Theories of theory of mind* (pp. 39–74). Cambridge, UK: Cambridge University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Perner, J., Gschaider, A., Kühberger, A., & Schrofner, S. (1999). Predicting others through simulation or by theory? A method to decide. *Mind and Language*, *14*, 57–79.

Perner, J., & Kühberger, A. (2002). Framing and the theory–simulation controversy: Predicting people's decisions. *Mind and Society*, *6*, 65–80.

Perner, J., & Kühberger, A. (2003). Putting philosophy to work by making simulation theory testable: The case of endowment. In C. Kanzian, J. Quitterer, & E. Rungaldier (Eds.), *Persons. An interdisciplinary approach* (pp. 101–116). Wien: öbv-hpt Verlagsgesellschaft.

Russell, B. (1919). On propositions: What they are and what they mean. *Proceedings of the Aristotelian Society, 2,* 1–43.

Schäfer, M. L. (1980). Das Problem der Intersubjektivität in der philosophischen Grundlegung der Psychriatrie. In G. Strube, *Die Psychologie des 20. Jahrhunderts, 20. Bd., Ergebnisse für die Medizin* (Vol. 2, pp. 63–77). Zürich: Kindler.

Scheler, M. (1973). Vom fremden Ich. In Gesammelte Werke VII: Wesen und Formen der Sympathie. In M. S. Frings (Ed.), *Die deutsche Philosophie der Gegenwart* (pp. 209–258). Bern, München: Francke.

Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? *Psychological Science*, *9*, 340–346.

Sellars, W. (1956). Empiricism and the philosophy of mind. In K. Gunderson (Ed.), *Minnesota studies in the philosophy of science* (Vol. 1, pp. 253–329). Minneapolis: University of Minnesota Press.

Stich, S., & Nichols, S. (1992). Folk psychology: Simulation or tacit theory? *Mind and Language, 7,* 35–71.

Stich, S., & Nichols, S. (1995). Folk psychology: Simulation or tacit theory? In M. Davis & T. Stone (Eds.), *Folk psychology* (pp. 123–158). Oxford, UK: Blackwell.

Stich, S., & Nichols, S. (1997). Cognitive penetrability, rationality and restricted simulation. *Mind and Language*, *12*, 297–326.

Strahilewitz, M. A., & Loewenstein, G. (1998). The effect of ownership history on the valuation of objects. *Journal of Consumer Research*, *25*, 276–289.

Thaler, R. H. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, *1*, 39–60.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Van Boven, L., Dunning, D., & Loewenstein, G. (2000). Egocentric empathy gaps between owners and buyers: Misperceptions of the endowment effect. *Journal of Personality and Social Psychology*, *79*, 66–76.

12

Why Self-Ascriptions Are Difficult and Develop Late

RADU J. BOGDAN

Many philosophers and a few psychologists think that we understand our own minds before we understand those of others. Most developmental psychologists think that children understand their own minds at about the same time they understand other minds, by using the same cognitive abilities. I disagree with both views. I think that children understand other minds before they understand their own. Their self-understanding depends on some cognitive abilities that develop later than, and independently of, the abilities involved in understanding other minds. This is the general theme of this chapter.

The argument focuses on what I take to be the core of understanding minds, namely, ascriptions of representational states or attitudes, such as desires or beliefs, whose relation (or directedness) to what they represent is registered in some fashion. I will dub this relation "representingness" (and treat it as equivalent to what philosophers call intentionality or aboutness). The argument does not apply to nonrepresentational states, such as feelings, or states whose representingness is not at stake (as when I notice that I begin to see in the dark) or is not registered at all. To save space, I will shorten representational attitudes to (simply) "attitudes" and ascriptions of attitudes to (simply) "ascriptions." I call the ascriptions of one's own attitudes "self-ascriptions" and those aimed at the attitudes of others "other-ascriptions." In what follows I will be concerned only with ascriptions that are sensitive to the representingness of the attitudes involved.

The argument is that self-ascriptions that register their own representingness are different, harder, and later than other-ascriptions because the former require abilities that (1) are more complex than those required by the latter and (2) develop after the age of 4, when basic other-ascriptions are already in place. The first section documents these claims. The section that follows explains in general terms why self-ascriptions are harder than other-ascriptions. The third section argues that current theories fail to account for these asymmetries. I then turn to autobiographical memory for clues to the nature of self-ascriptions. The final section sketches a neuropsychological account of self-regulation and self-metarepresentation that explains why self-ascriptions develop later than other-ascriptions.

ASYMMETRIES

The data presented below are rather sketchy and not uncontested but, if read as I suggest, they seem to favor the developmental and cognitive asymmetry I am proposing. First, there is some *direct* evidence of a developmental asymmetry. Children understand the desires of others around the age of 2 (Wellman, 1990) but do not grasp their past unfulfilled or changed desires even a year later (Gopnik, 1993). The same seems to be true of false belief: estimates by Gopnik and Astington (1988) place the self-ascription of false belief in the 4- to 5-year-old interval but its other-ascription in the 3- to 4-year-old interval (see also Flavell, Green, & Flavell, 1995, pp. 53–54, for a survey).

The *indirect* evidence is more robust and points in the same direction. The child's *metacognition,* required for an explicit awareness of self-attitudes, develops after age 5 and takes time to mature; *introspection* is estimated to emerge several years later (Flavell et al., 1995; Nelson, 1996). Before ages 7–8 children often fail to identify their own past thoughts and their contents, even when recent (Flavell et al., 1995, pp. 80–81). If young children master self-ascriptions as soon as they master other-ascriptions and if, as seems likely, they master the basics of the latter around 4, what explains these delays in metacognition, introspection and thought identification? Couldn't it be because self-ascriptions develop late and slowly? Self-ascriptions, particularly of past attitudes, often require the *inhibition* of one's current cognition. Inhibition develops only after the 3–4 period (Bjorklund, 2000; Houdé, 1995; Leslie, 2000).

There is also neurological evidence for asymmetry. The early theory of mind involved in face and gaze recognition and the representation of agency is done mostly in the *left* hemisphere (Baron-Cohen, 1995). The left hemisphere excels at selecting and processing a single, dominant mode of representation, and blocking out all the others—which is how the other-ascriptions of the first 3 years' work. In contrast, most of the later theory-of-mind work, including self-ascriptions, is done by the *right* hemisphere, in exchanges of information with the left hemisphere. The clinical evidence lends its support, too. Damage to the left hemisphere, manifested in autism, compromises the ability to handle joint attention and other-ascriptions of false belief—all failures of early theory of mind. In contrast, damage to the right hemisphere—in schizophrenia—prevents complex self- and other-ascriptions while leaving intact simple other-ascriptions (Corcoran, 2000; Frith, 1992).

A most telling evidence for the asymmetry thesis is that, until age 4 to 5, children lack *autobiographical memory* (Conway, 2002; Nelson, 1996). This sort of memory, which develops gradually, is needed in self-ascriptions of past attitudes. As far as I can tell, most experiments with and analyses of self-ascriptions of past desires and beliefs are about the *immediate* past (several minutes or hours). What the child has to do (which is not easy before 4) is inhibit the content of a current desire or belief and recall a past content. This is not enough for testing self-ascriptions that are sensitive to their representational relations. The memory of a past content is not the memory of a *representational relation* to that content.

It is the representational relations or representingness of one's own attitudes—which I will call *self-representingness*—that poses the mightiest challenge to self-ascriptions, delaying and lengthening its development, and making it harder to acquire and use than other-ascriptions. The notion of a sense of self-representingness is broad and allows for various resources—procedures, images, schemes, thoughts—that can do the job. The next section ventures a first explanation of why self-representingness is hard to register.

THE ELUSIVENESS OF SELF-REPRESENTINGNESS

It helps to begin by making explicit what it takes to self-ascribe an attitude. I present only three conditions that matter to our discussion:

1. Evidential basis: having evidence for self-ascribing an attitude, in the form of inner experiences, mental activity, introspection, inference, etc.

2. Right concepts: possessing the appropriate concepts of attitudes, such as desire or belief.
3. A sense of mental self-representingness: understanding one's own mental states as representational, as being about something (there is also a sense of bodily and behavioral self-directedness to a target).

The literature on early theory of mind often treats inner experiences of mental states and their contents as sufficient for self-ascriptions. This is not enough. Consider one's own desire. Although a desire is directed at something, to recognize having the desire, through some inner experience, is not to recognize *its representingness*. A child may recognize his or her desire from how it is experienced inside and from its content (what is desired) *without* recognizing the relation between experience and content. Many observations and experiments misdiagnose self-ascriptions by measuring only the inner experience of an attitude and some experience-based concept but missing the representingness of the attitude. Such misdiagnoses may actually suggest that young children manage self-ascriptions at the same age as, or even earlier than, they manage other-ascriptions.

Here is an analogy that underscores the importance of having a sense of representingness. Many organisms learn words or signs by *associating* a visual or sound experience with an object or event (as content or referent) without recognizing the referential *relation* between the experience and its content. This is the key difference between a parrot and a human child. The latter (but not the former) recognizes the relation in question because she is a naive psychologist who can grasp the reference relations between other minds and objects and events (Tomasello, 1999). If this recognition is not factored into the analysis, it looks like word acquisition is a widespread animal trait—just as self-ascription looks easy without the recognition of self-representingness. An experience of a mental state merely associated with its content is not good enough for the mastery of word meanings or self-ascriptions.

I will center my analysis on the self-ascriptions of past false belief. It is important to focus on self-ascriptions of past and mismatched attitudes, such as unfulfilled desires or false beliefs, rather than on self-ascriptions of current and successful attitudes, because only awareness of mismatch measures one's sense of self-representingness. One knows that one mentally aimed at something when one knows that one missed it. This is why false belief tests the possession of the concept of belief. The pastness of an attitude forces a mental reconstruction that separates self-representingness from a current experience of an attitude or its current content. The self-ascription of a past false belief illustrates this point best because, unlike one's own past desires or emotions, one's own past beliefs are more abstract and carry fewer experiential traces, so to speak. Whereas one can almost relive past desires or past emotions, one has to *think* (often hard) about past false beliefs.

In the false-belief experiments, the child needs to do at least the following to figure out the false belief of *another* person, whom I call target:

A1. Inhibit her propensity to treat what she currently knows about the facts of the case as the target's actual belief.
B1. Recover a (visual) memory of the target's past relation to a content and represent that relation as the target's actual (and false) belief.

In the experiments on self-ascription (such as the Smarties box that turns out to contain pencils), to figure out *her own* past false belief, the child must:

A2. Inhibit her propensity to treat her current representation of the content (pencils) as her earlier belief.
B2. Recover a memory of an earlier representation of the content (Smarties) and represent it as her earlier (and false) belief.

There is an *executive* symmetry here: the first conditions, A1 and A2, call for inhibition—a crucial development around age 3–4, widely credited with enabling the ascriptions of false belief (Harris, 1992; Leslie, 2000; Perner, 1991; see also Bogdan, 2003). The asymmetry between the two kinds of ascriptions is *cognitive* and concerns the representingness of the attitudes ascribed. Think of it intuitively: the child has no *perceptual* clues to the representingness of her own beliefs, that is, to the fact that her beliefs are representing (the relation) their contents; but she has such clues to the representiational relation of a third-person belief. In A1 the child sees not only what the target perceives and therefore believes but also sees that the target has a perceptual and hence belief *relation* to the world. (As noted below, the young child's concept of belief seems to be modeled on that of perception.) But in A2 all the self-ascriber perceives are pencils (the content); she does not perceive the *relation* to that content. In B1 the child can recall visual memories about a past attitude of the target in relation to the content, whereas in B2 the self-ascriber cannot recall any clues pointing to her belief as a *relational* attitude; all she remembers is a content—the Smarties box. The representingness of her belief is not evident in her memory. And it could not be, because, before age 4, the child does not yet have an autobiographical memory that could retrieve her own past attitudes *as* representational relations to contents. The memories of early years are exclusively semantic and episodic (Conway, 2002; Nelson, 1996), and they deliver only contents experienced in some form. I develop these points later. Right now I want to draw an implication from what has been said so far.

We should not underestimate the role of *visual* evidence in the development of ascriptions. Several decades of research have shown that the child's (and possibly the ape's) understanding of other minds begins and develops through visual observation of others—their faces and facial expressions, looks, gaze, bodily posture, movement, and behavior—and of the contexts of social interaction. Congenitally blind children take longer to develop an understanding of other minds (Hobson, 1993). The more complex and invisible attitudes, such as belief and intention, may be modeled on the simpler and more visible ones, such as perception and desire (Wellman, 1990). It is likely that the early sense that children have of the representingness of other-attitudes emerges out of their visual recognition of gaze and attention (Moore & Dunham, 1993; Tomasello, 1999). The point I am making is not that a theory of mind develops out of the visual observation of others. Many species engage in such observation, but very few do theory of mind. The point, rather, is that visual evidence is important for activating, maturing, and guiding the theory-of-mind abilities to detect other-attitudes.

There is no corresponding visual evidence for detecting *self-*representingness. Children represent the contents of their own attitudes and sense their presence in the mind from internal signals but cannot observe their representingness as relations to contents. People do not *observe* their perception or belief *as* relations to the world, particularly in the case of past such relations. Yet, children develop nonvisual means of registering self-representingness from *inside* their minds. How do they do it? A brief survey of current theories does not yield a satisfactory answer to this question.

OTHER VIEWS

The main accounts of self-ascriptions are the theory-theory view and the simulation view. I locate my own view relative to them by reversing the predictions of failure that each makes against the other. The theory-theory predicts that simulation theory, by positing self-ascriptions to be prior to and easier than other-ascriptions, would be distressed by evidence of temporal and cognitive symmetries between the two sorts of ascriptions. Simulation theory predicts that theory-theory, by positing such symmetries, would be distressed by evidence of asymmetries favoring self-ascriptions as earlier and easier. Against the theory-theory, my analysis proposes a temporal and cognitive asymmetry, but, against simulation theory, it finds other-ascriptions earlier and easier than self-ascriptions.

The *theory-theory* view is committed to the developmental and conceptual symmetry between other- and self-ascriptions (Gopnik, 1993; Gopnik & Meltzoff, 1997; Perner, 1991). I take the general case made for asymmetry in the first section to count against the theory-theory position. But I also have more specific arguments. The theory-theorist may agree that there are differences between self- and other-ascriptions with respect to evidence and cognitive resources, such as inhibition and memory, but insists that they need not entail a difference in the *concepts* utilized (Gopnik, 1993; Gopnik & Wellman, 1992; Malle, personal communication; Perner, 1991). And sameness of concepts appears to invalidate the asymmetry thesis. Actually, it doesn't, as I argue next.

First, my asymmetry hypothesis is that, helped by visual evidence, young children first apply the concepts of attitudes to others, and need more time, more cognitive effort, different evidence, and new abilities to figure out how to apply the (alleged) same concepts to themselves. Same concepts notwithstanding, it is still the case that self-ascriptions can be harder and develop later than other-ascriptions. A second argument points to evolution. The asymmetry is not just a matter of evidence and cognitive resources. It is also a matter of the *function* of a theory of mind. A good case can be made that in humans and possibly other primates a naive theory of mind first evolved to deal with conspecifics, not with selves. The most pressing challenges posed by conspecifics are in ongoing contexts of social interaction and involve observable features of conspecifics, such as facial expression, gaze, bodily posture, behavior, communication, and the like (Bogdan, 1997; Tomasello, 1999; Whiten, 1991). It is unlikely that the basic concepts and ascription strategies of theory of mind evolved for reasons other than registering visible relations between conspecifics and the world, ascriber and conspecifics, and among conspecifics themselves. Looked at in this light, the turn to self-ascription is a late evolutionary and developmental event, and not a constant counterpart to other-ascriptions.

The argument from evolution and the ones that follow begin to challenge the same-concepts claim. Suppose that young children are primarily interested in other people and that they first ascribe to them *observable* attitudes, such as seeing. Suppose also that these ascriptions are directed at *concrete*, *spatiotemporally* defined items in the world, such as objects and events, and not at propositions, as they will be later. Suppose, finally, that these early other-ascriptions are *egocentric*, in that they reflect the child's ongoing motivation and perception, and *situated* because they are tied to current contexts of interaction with others. I have defended this characterization of early theory of mind elsewhere (Bogdan, 1997, 2000, 2003), so I will move to the relevant arguments it entails.

One argument is this: theory-theorists tend to think that (unless innate or essentialist) concepts are normally formed and revised *in response to* perceptually accessible facts and regularities in their domains (Gopnik, 1993; Gopnik & Meltzoff, 1997; Perner, 1991; Wellman, 1990). If that is so, then it is hard to see how other-concepts could *initially* be the same as self-concepts, when the facts, regularities, and the perceptual evidence revealing them in the other-domain are so different from those in the self-domain in key respects, particularly concerning self-representingness. This initial difference does not preclude a later alignment of self- to other-concepts, when the older child acquires a sense of his or her own mental representingness.

Another argument is the following: theory-theorists agree that early theory of mind is not *metarepresentational*, for it does not represent the rather abstract and invisible representational relations of complex attitudes, such as opinions, intentions, or hopes. However, one's own past attitudes, particularly false or unfulfilled ones, also have a rather abstract and invisible representingness. The early concepts for other-ascriptions are not equipped to handle this sort of past self-representingness—and I do not think they do. After the metarepresentational turn around age 4, the child acquires a new supply of concepts that track self-representingness. The question is whether the new concepts are self-other symmetrical or not. It is a question I will tackle in the final section. The answer is less simple than expected.

I turn next to the *simulation* view, with its two versions. The practical-reasoning version of Robert Gordon (1986), which does not require ascriptional concepts, has two options. One is an ascent routine that habituates the child to the link between a content in mind and first-person locutions, such as "I believe that . . . " I do not see how this works for "I falsely believe that . . . " and particularly for "I believed falsely that . . ." without begging the question at stake, which is how one's own past representingness is represented. As Gordon (1993, p. 45) notes, the ascent routine cannot deliver it. The routine provides at best a head start for counterfactual and imaginative simulation, which is the other option. How would this option work?

According to Paul Harris (1992), in the test of one's own past false belief, the child must imagine the proposition she originally entertained and took to be true (Smarties in the box) and, inhibiting her current knowledge (of pencils), report on its usual contents (Smarties). Before 4, children have the same difficulty representing their own past false beliefs as well as those of others, for they do not have enough counterfactual imagination. But how would that imagination track self-representingness? Young children imagine mostly in visual terms, and those, I argued earlier, are not the right terms for one's own *past* representingness.

How about introspective simulation based on ascriptional concepts (Goldman, 1993; Harris, 1992)? This view advocates an asymmetry in the opposite direction from mine: self-ascriptions develop earlier and are easier than other-ascriptions. One problem is that the introspection that classifies attitudes need not reach beyond experience, current or recalled. As Alvin Goldman notes (1993, p. 105), introspection can identify the type and content of an attitude but not its representingness and hence truth value. Another problem is that a representingness-sensitive introspection may be a late development, perhaps as late as age 7 or 8 (Flavell et al., 1995), which is later than the onset of self-ascriptions. Like Harris, Goldman (1993, p. 43) thinks that the key obstacle for a young introspector attempting to recognize her own past false belief is sorting out and dating conflicting representations. This is right, but even with representations of current and past contents sorted out and dated, the young introspector still has no idea of her former self being representationally related to and actually misrepresenting a past content. She just recalls a past content that is different from a current one.

Theory-theory and simulation are not the only accounts of self-ascriptions. There are modular accounts, such as those of Leslie (2000) and Baron-Cohen (1995), and also accounts of forms of self-other coordination, such as those of Gopnik and Meltzoff (1997) and Barresi and Moore (1996). And the list is not complete. I do not have the space to discuss them, but I will say this much, without argument: as survival devices, modules evolved to deal with other-attitudes, not one's own. Their sensitivity to faces, eye, gaze, and bodily signals is evidence of sensitivity to others and their representational relations to the world. Nichols and Stich (2003) posit an innate self-monitoring module, active since age 2, which detects one's own mental states. The problem is that detection (like inner experience and introspection) may signal the presence and partly the type of an attitude but not its representational relation. Coordination mechanisms map one's experiences onto the attitudes of others, not of selves, and such mappings need not be sensitive to representational relations. In general, modular and coordination accounts best explain the other-ascriptions of early childhood but not later self-ascriptions (Bogdan, 1997, 2000).

Brief and fast-paced as my critical survey has been, it should be recalled that it concerns solely views about self-ascriptions that are sensitive to the representational relations of attitudes (the hard ones) and not other sorts of self-ascriptions—say, of feelings or of current or immediately past mental states recognized solely through their internal experiences or their contents (the easy ones). Such experiences and the partial concepts based on them are necessary, as early precursors, for the development of self-ascriptions—a necessity explained in various ways by the views just surveyed. But their story remains incomplete because a sense of self-representingness is not in the picture.

If the early theory of mind is insensitive to self-representingness, what are the resources that generate that sensitivity, and why do they develop only after age 4? I propose to look for clues to an answer in the domain of memory, because it plays a major role in past self-ascriptions and because its development parallels in important respects that of theory of mind.

MEMORIES OF PAST ATTITUDES

We have immediate access only to our current attitudes. Past attitudes must be retrieved from memory. The question is what sort of memory could do this job. Young children have good memory for things, events, and situations. This is *semantic* memory. It retrieves only content, without its actual experience. Another sort of memory, called *episodic*, retrieves experiential details of past contents, linked to spatiotemporal contexts, and vivid reactions and emotions. Episodic memory also operates in young children and perhaps some nonhuman species (Clayton, Griffiths, Emery, & Dickenson, 2002). Episodic memories are represented in the same brain areas as are actual experiences (Conway, 2002). This is significant because it suggests that, like ongoing mentation, episodic memory accesses only past experiences or their contents but not past self-representingness. Furthermore, the reliving of past experiences or contents may re-create a vivid sense of having seen or desired something but not of having believed something, which is a much less vivid attitude. This is why one's own past perceptions or desires are easier to recall than past beliefs.

An objection raised by Bertram Malle (personal communication) touches on both episodic memory and the central issue of this chapter. Why should a self-ascriber need a *separate* sense of self-representingness? Wouldn't it be enough to reexperience episodically one's desire (say) for coffee this morning, which may combine memories tagged with the time it happened, who the self was, what was desired, and the sort of mental state it was? And doesn't one do the same with self-ascriptions of a current attitude—join an experience of the type of attitude with an experience of its content and of the self that has these experiences? Aren't young children capable of all these exploits, both in episodic memory and current self-ascriptions?

Suppose they are. It is still a developmental fact that children recall things episodically very early but cannot do self-ascriptions that are sensitive to their representingness until a few years later. It is also a developmental fact, proven by infantile amnesia, that the young children's episodic memories evoke relatively short-lived experiences, and that access to such memories tends to degrade rather quickly (Conway, 2002). The reason, I think, most experiments with young children's self-ascriptions appeal only to a *recent* past is that they test only their episodic memory. If older children and adults have long-term memories, it can't be solely because they have episodic memory. I suspect that long-term memories have something to do with being able to represent one's past attitudes and their representingness. So, what sort of memory could do it?

The answer is *autobiographical* memory. It is the sort of memory that children lack until around age 3 and halfway to 4, and that develops gradually until 6 (Bjorklund, 2000, p. 264; Nelson, 1996, pp. 157 and 162), which is also the interval when representingness-sensitive self-ascriptions develop. Autobiographical memory terminates infantile amnesia by integrating and consolidating episodic memories in autobiographical terms and enabling a retrieval of past attitudes. How does this work? Autobiographical memory is said to add to its episodic basis an autonoetic or quasi-introspective consciousness and re-creative thinking. These new abilities seem uniquely human and develop in the 4–6 age interval (Conway, 2002). I schematize the ontogenesis of autobiographical memory as follows:

> Semantic memory + [Re-creation of experiences in terms of perceptual vividness, spatiotemporal framing, and affective associations] =
> Episodic memory
>
> Episodic memory + [Autonoetic consciousness + Re-creative thinking] =
> Autobiographical memory?

Yet, reliving past experiences consciously and recreatively may still not be enough for autobiographical memory. One may recall episodically, through imagery or inference, the passing show of past events without representing one's past attitudes as true or false or referring to this or that. What else is needed? Josef Perner's answer (1991, pp. 163–169; 2000) is *metarepresentation*. It explains why episodic memory turns autobiographical and why (on my analysis) the latter can represent one's past self in various representational relations to things and situations.

This idea is plausible for two reasons. In remembering autobiographically past attitudes, we *represent* ourselves back then *representing* the contents of past attitudes. This is what metarepresentation does. The idea is also plausible because, on most accounts, metarepresentation develops around age 4. The sort of metarepresentation Perner has in mind is symmetrically shared by self- and other-ascriptions. For reasons discussed earlier, it does not seem quite the right sort. If autobiographical memory is required for self-ascriptions of past attitudes, we need to look for another sort of metarepresentation, one that is *intrinsically* sensitive to self, as is autobiographical memory itself. This is the issue I turn to next and last.

MINDING OUR OWN MINDS

My proposal is that a sense of self-representingness grows out of the *executive* tasks of self-regulation of the new mental activities that develop after the age of 4. The main self-regulatory tasks consist in holding in mind many representations in an active state for an extended period, monitoring, controlling, integrating, and manipulating the information needed for a task, and inhibiting task-irrelevant information. Most of this work can only be done in terms of what and how one's thoughts represent what they do. This is why this *intramental* work calls for a sense of one's own thoughts being related to what they represent—a sense of mental self-representingness. The demonstration of this thesis can only be sketched here in a few telegraphic steps.

I begin by contrasting two metaphors. Until the 3–4 age interval the young mind operates on a single central screen, where perceptual and memory inputs are displayed and constantly updated by new inputs. It is a mind largely, though not entirely, confined to current motivation and perception. The young mind can imagine beyond the current inputs but still within their frame and theme. Think of the imaginative stance of young childhood as a sort of little screen or box that opens in a corner of, and from inside, the larger screen dominated by current perception and/or memory. This is a simplification, of course, but the contrast it highlights is real.

After age 3, the young mind is shaken by several mental commotions, executive as well as cognitive, and revolutionary in their cumulative impact. Chief among them are the inhibition of current perception, the linguistic recoding and representational explicitation of earlier procedural competencies, the development of short-term memory as the workspace where multiple and alternative representations can be maintained, manipulated, and integrated in various formats (Diamond, 2001; Houdé, 1995; Karmiloff-Smith, 1992). These developments liberate the young mind from the captivity of single-screen or uniplex mentation and enable it to entertain simultaneously, in different but interconnected mental screens, nested sets of alternative and often conflicting representations of actual and nonactual, current, past, and counterfactual situations. A multiscreen or multiplex mentation comes of age. It creates its own pressures for internal self-regulation in the form of supervisory capacities operating explicitly on and with thoughts in terms of their representational relations and features (Perner, 1998; Shallice & Burgess, 1993). The child's mind thus develops an internal metarepresenter, or *metamind*. The chief neural platform of this new metamind is the (dorsolateral) *prefrontal cortex*, with the *integrative connectivity* handled mostly by its right hemisphere and reaching across large regions of the brain. The growth of this platform is most dramatic in the 3–6 age interval (Diamond, 2001).

Nothing in the story so far mentions theory of mind. The self-regulatory job of the metamind is the basic phenomenon. It is quite a different question whether, in order to carry out its self-regulatory functions, the metamind develops its own metarepresentational tools or recruits those of the child's theory of mind. Similar or nearby brain structures, which develop in the 4–7 interval, seem to have a hand in many executive and theory-of-mind tasks. Such neural and temporal proximity and the idea of control by metarepresentation may suggest that it is developments in theory of mind that are co-opted for intramental self-regulation (Perner, 1998; to some extent, Frith, 1992).

This scenario is possible but unlikely, I think, because of the very nature of what is to be monitored and controlled, and how. Consider intention. It is hard to see how metarepresenting one's own intentions, for self-regulation, could result simply from recruiting a theory-of-mind concept. Intentions are recognized and monitored internally in order (among other things) to distinguish between actions caused by our desires and plans and those reacting to external events. Failure to make this distinction impairs the act of intending and may result in delusions of control and other passivity experiences (Frith, 1992). Failure to make the same distinction in the case of other people is unlikely to have similar effects. One must first have an internal sense of what it is to monitor and control one's own acts and representations, whether sensorimotor or mental, before one can conceptualize such acts and representations.

There is also a neurological reason for doubting that developments in theory of mind are responsible for self-metarepresentation. Most of the latter work is done by the right hemisphere in exchanges with the left hemisphere, whereas early other-ascriptions activate mostly the task- and domain-specific left hemisphere (Brownell, Griffin, Winner, Friedman, & Happé, 2000). And, as noted in the first section, damage to the left and frontal brain affects other-ascriptions, whereas damage to the right hemisphere impairs only higher-order ascriptions and self-metarepresentation (Corcoran, 2000).

Fortified by these reasons, I propose that between ages 4 and 7, the self-regulatory metamind develops its own *predispositions* for self-metarepresentation and thus triggers the development of a sense of self-representingness.

To get a better handle on this proposal and see its cerebral plausibility, consider the following (much simplified) analogy. To monitor and control its motor actions, an organism must have *metamotor* information. Suppose its actions are represented by motor images that track bodily positions relative to visual stimuli from the targets of its actions. If the organism is just reactive or on automatic pilot, the motor images are fed into preset action schemas, and all is well. But if the organism is endowed with top-down control and attention, and needs to initiate a new action or modify an action or watch closely what it is doing, it must be able to track the motor images themselves *as they relate to their targets*.

Tracking the representational relations of motor images is the job of *metamotor* images as second-order motor representations. Whereas first-order motor images represent actions relative to bodily states and external targets, the metamotor images compare what first-order motor images represent with internal models from motor memory and with action predictions made by a planning center (Damasio, 1999; Jeannerod, 1997). The metamotor comparisons between first-order motor images and memories and planning predictions enable an organism to register and control the representational relations of its motor images because the results of the comparisons provide information about the organism being *related to a target* and about whether it is on target, properly directed at it or not, and if not, by how much. Metamotor images provide a sense of motor self-representingness because they enable an organism to do things with and to motor images *in terms of their representational relations*.

I suggest we think in the same control-of-activity spirit about the self-regulatory work of the metamind. Multiplex mentation is a new domain of activity to be mapped and supervised. It happens to be an intramental domain, inhabited by one's own thoughts and thought processes. Given that thoughts represent all sorts of targets (worldly, mental, abstract), and cause as well as get feedback from other thoughts in terms of what they represent, the self-regulatory work of the metamind must be *metarepresentational* and engage thoughts at their representational joints, such as reference, coherence, and truth value. As a result, the metamind generates a sense of one's own thoughts being *related* to what they represent, which is a *mental* sense of self-representingness.

So construed, self-metarepresentation may have a generic format that treats one's own thoughts as mental states that represent and have internally recognizable functions (to remember, to infer, to act on, etc.) but are not yet classified *as* desires, beliefs, intentions, and so on, according to a theory of mind. It is when one's metamind must represent one's own thoughts *as theory-of-mind attitudes* that this generic format may become explicitly structured by theory-of-mind concepts. Some of these concepts probably build on their precursors, particular on the internal symptoms of desires, beliefs, and so on. It is equally possible that the executive demands on the emerging metamind may force a dramatic revision in the child's earlier theory of mind, if the latter is to integrate self- and other-ascriptions in ways that have self-regulatory impact. After all, developments after age 3–4 acquaint and confront the child's mind with an entirely new domain—his or her own thoughts now recognized explicitly as representational. The child's theory of mind must adapt to this new domain and reconcile it with the domain of other minds. An outcome of this process may be a new conceptual cartography of the mind, integrating self- and other-ascriptional concepts. But the point I am emphasizing here is that the *initial* self-metarepresentation is likely to result from intramental self-regulation.

To return to our parallel topic for a last show of support, an internally driven and attitude-free self-metarepresentation seems also to be involved in autobiographical memory. One's episodic memories beam back vivid snippets of an original experience, testifying to its authenticity. What confers a sense of self-representingness and veracity to those memories is the monitoring, integration, and evaluation of the representations involved in terms of how they fit together, how they organize the information thematically and narratively, and so on. Autobiographical memories convey a sense of their representingness *without* necessarily representing self-attitudes in a theory-of-mind format. One need not have a past belief about an event in order to remember the event autobiographically. It is the other way around. One remembers autobiographically the event because of how the work of one's memory meshes with that of one's metamind. The resulting memories in turn allow the recovery of past attitudes toward the remembered event.

Time to sum up. If we ask why self-ascriptions are later and harder than other-ascriptions, the answer I have proposed is that, unlike the latter, the former are grounded in an internally driven self-metarepresentation. Whereas many of the concepts and schemes of other-ascriptions emerge early in the child's theory of mind, the self-metarepresentation required for self-ascriptions develops only after the age of 4, for neuropsychological reasons having to do with brain development and self-regulation rather than theory of mind.

ACKNOWLEDGMENTS

My warm thanks to Janet Astington, Sara Hodges, Carolyn Morillo, and Josef Perner for helpful comments; to audiences at the University of Oregon, Universidad Autonoma de Barcelona, and University of Bucharest for their reactions to early versions of this chapter; and particularly to Bertram Malle for detailed observations and excellent suggestions.

REFERENCES

Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, MA: MIT Press.

Barresi, J., & Moore, C. (1996). Intentional relations and social understanding. *Behavioral and Brain Sciences, 19*, 107–122.

Bjorklund, D. F. (2000). *Children's thinking*. Belmont, MA: Wadworth.

Bogdan, R. (1997). *Interpreting minds*. Cambridge, MA: MIT Press.

Bogdan, R. (2000). *Minding minds*. Cambridge, MA: MIT Press.

Bogdan, R. (2003). Watch your metastep: The first-order limits of early intentional attributions. In C. Kanzian, J. Quitterer, & E. Runggaldier (Eds.), *Persons*. Vienna: obv & hpt.

Brownell, H., Griffin, R., Winner, E., Friedman, O., & Happé, F. (2000). Cerebral lateralization and theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience*. New York: Oxford University Press.

Clayton, N. S., Griffiths, D. P., Emery, N. J., & Dickenson, A. (2002). Elements of episodic-like memory in animals. In A. Baddeley, J. P. Aggleton, & M. A. Conway (Eds.), *Episodic memory: New directions in research* (pp. 232–248). London: Oxford University Press.

Conway, M. A. (2002). Sensory-perceptual episodic memory and its context: Autobiographical memory. In A. Baddeley, J. P. Aggleton, & M. A. Conway (Eds.), *Episodic memory: New directions in research* (pp. 53–70). London: Oxford University Press.

Corcoran, R. (2000). Theory of mind in other clinical conditions. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience*. New York: Oxford University Press.

Damasio, A. (1999). *The feeling of what happens*. New York: Harcourt, Brace.

Diamond, A. (2001). Normal developments of prefrontal cortex from birth to young adulthood. In D. T. Stuss & R. T. Knight (Eds.), *The frontal lobes* (pp. 466–503). Oxford, UK: Oxford University Press.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1995). Young children's knowledge about thinking, *Monographs of the Society for Research in Child Development*, *60*(243), 1–96.

Frith, C. (1992). *The neurological basis of schizophrenia*. Hillsdale, NJ: Erlbaum.

Goldman, A. (1993). The psychology of folk psychology. *Behavioral and Brain Sciences*, *16*, 15–28.

Gopnik, A. (1993). How we know our minds. *Behavioral and Brain Sciences, 16*, 1–14.

Gopnik, A., & Astington, J. (1988). Children's understanding of representational change. *Child Development*, *59*, 26–37.

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

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

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

Gordon, R. (1993). Self-ascriptions of belief and desire. *Behavioral and Brain Sciences, 16*, 45–46.

Harris, P. (1992). From simulation to folk psychology. *Mind and Language, 7*, 120–144.

Hobson, R. P. (1993). *Autism and the development of mind*. Hillsdale, NJ: Erlbaum.

Houdé, O. (1995). *Rationalité, développement et inhibition*. Paris: Presses Universitaires de France.

Jeannerod, M. (1997). *The cognitive neuroscience of action*. Oxford, UK: Blackwell.

Karmiloff-Smith, A. (1992). *Beyond modularity.* Cambridge, MA: MIT Press.

Leslie, A. (2000). Theory of mind as a mechanism of selective attention. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 1235–1247). Cambridge, MA: MIT Press.

Moore, C., & Dunham, P. J. (Eds.). (1995). *Joint attention*. Hillsdale, NJ: Erlbaum.

Nelson, K. (1996). *Language in cognitive development*. Cambridge, UK: Cambridge University Press.

Nichols, S., & Stich, S. (2003). *Mindreading*. Oxford, UK: Oxford University Press.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Perner, J. (1998). The meta-intentional nature of executive functions and theory of mind. In P. Carruthers & J. Boucher (Eds.), *Language and thought*. Cambridge, UK: Cambridge University Press.

Perner, J. (2000). Memory and theory of mind. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 297–312). Oxford, UK: Oxford University Press.

Shallice, T., & Burgess, P. (1993). Supervisory control of action and thought selection. In A. Baddeley & L. Weiskrantz (Eds.), *Attention, selection, awareness and control* (pp. 171–187). Oxford, UK: Oxford University Press.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Wellman, H. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Whiten, A. (Ed.). (1991). *Natural theories of mind*. Oxford, UK: Blackwell.

PART IV

Language and Other Minds

This page intentionally left blank

13

Language as the Route into Other Minds

JANET WILDE ASTINGTON
EVA FILIPPOVA

He was about four, I think . . . it was so long ago.
In a garden; he'd done some damage
behind a bright screen of sweet-peas
—snapped a stalk, a stake, I don't recall,
but the grandmother came, and saw, and asked him:
"Did you do that?"

Now, if she'd said *why* did you do that,
he'd never have denied it. She showed him
he had a choice. I could see in his face
the new sense, the possible. That word and deed
need not match, that you could say the world
different, to suit you.

When he said "No," I swear it was as moving
as the first time a baby's fist clenches
on a finger, as momentous as the first
taste of fruit. I could feel his eyes looking
through a new window, at a world whose form
and colour weren't fixed.

but fluid, that poured like a snake, trembled
around the edges like northern lights, shape-shifted
at the spell of a voice. I could sense him filling
like a glass; hear the unreal sea in his ears.
This is how to make songs, create men, paint pictures,
tell a story.

> I think I made up the screen of sweet-peas.
> Maybe they were beans; maybe there was no screen:
> it just felt as if there should be, somehow.
> And he was my — no, I don't need to tell that.
> I know I made up the screen. And I recall very well
> what he had done.
>
> —PUGH, "The Beautiful Lie" (2002, p. 7)

Most of us will recognize the boy in Sheenagh Pugh's poem, whether our own child or grandchild, brother, nephew, or just a boy from the neighborhood. And many of us might identify with the poet as she recalls the magic of the precious moment of revelation. In fact, few moments in life are as eye-opening as young children's recognition of the possible—their realization that the world is not fixed, that the mind can entertain more beyond what is directly seen, heard, or touched. The poem dramatizes the suddenness of this realization. More prosaically, developmental psychologists study its slow dawning over the preschool years. Even so, both poet and psychologist acknowledge its enormous significance.

In this chapter we argue that young children's entry into the world of possibility is the crucial route into other minds. Moreover, it is a route that opens up only as their language abilities develop. The poem beautifully captures how language mediates our experience of the world—how it can be used to *re*-present events so that what is communicated need not match what is (or was) in the world. As children acquire language, they acquire the ability to think about possible, re-presented scenarios, and they become able to imagine what other people might think, want, or feel. A new world opens up for them as they realize that people's minds are private from one another and separate from the real world. They acquire a new sense of the possible: "That word and deed need not match, that you could say the world different, to suit you" (Pugh, 2002, p. 7). In this way they become aware of other minds or, in the terminology most commonly used in this research area, they acquire a "theory of mind."

THEORY OF MIND

In the most precise use, theory of mind refers to a domain-specific, psychologically real structure composed of an integrated set of mental state concepts used to explain and predict people's actions and interactions that is reorganized over time when faced with counterevidence to its predictions (Gopnik & Wellman, 1994). However, in recent years the term has been used more loosely to refer to a wide variety of abilities

(Astington & Baird, 2005). Sometimes it is used to refer broadly to social understanding in general. Other times it is used narrowly to refer to success on a specific task—the false-belief task—that assesses a child's ability to judge what somebody will do or say in a situation where their belief is different from what the child knows to be true. In a parallel fashion, while some researchers use the term to refer to development that begins with 9-month-olds' communicative abilities, others restrict their use of it to designate the 4-year-olds' specialized ability to reflect on (that is, to metarepresent) others' beliefs. Despite some controversy surrounding this diversity of usage, the term has become widely accepted and could not easily be replaced now.

We use the term here to refer to young children's developing abilities to understand self and others as agents who act on the basis of their mental states (i.e., beliefs, desires, emotions, intentions). In our view, theory of mind is a multifaceted system, grounded in social perception in infancy, that leads to the awareness of mental states and perspective-taking abilities that develop in the preschool and early school-age years. Theory of mind enables children to explain, predict, and interpret human behavior, where "behavior" is not only physical actions but also speech acts. That is, theory of mind underlies the ability to interpret human action and communication. It is the key to understanding other minds, and, crucially, language plays a pivotal role in its development.

LANGUAGE

Like theory of mind, language too is a complex, multifaceted construct. Human language, unlike communication systems used by other species, has the potential for communicating about things outside the immediate context—about the past and future, and about hypothetical possibilities that are represented in the mind. The linguistic system thus serves two purposes: communication and representation. Although other species communicate and have mental representations, only humans use one and the same system for both purposes, that is, social communication and verbal, mental representation. We argue that it is because the same system is used for both these purposes that language can provide a route into other minds.

Language competence involves progressively acquired skills in different subdomains of the linguistic system. On the one hand, language competence involves knowledge of different aspects of linguistic form (*phonology, morphology,* and *syntax). On the other hand, it involves knowledge of lexical and discourse meaning (*semantics*). In addition, language competence entails an ability to express and interpret meanings in communicative exchanges *(pragmatics*). All these components of language as a system are gradually acquired by each typically developing child whose own maturing cognitive capacities allow for this achievement to take place. Importantly though, language acquisition takes place in a specific context and under specific circumstances, all of which influence the language competency a child achieves. It is in the midst of lively social interaction and everyday discourse that a child becomes a competent member of a larger social group. Hence, although language acquisition depends on individual cognitive capacities, it also depends, to a large extent, on the social and cultural context.

For this reason, when considering the role of language as a route into other minds, an important distinction needs to be made between *intraindividual* and *interindividual* aspects of language. The intraindividual aspects of language encompass the child's own linguistic abilities, including his or her syntactic, semantic, and pragmatic skills. The interindividual aspects of language are the social or environmental features of the context, including all the communicative interactions a child is engaged in. As such, human language is used both as an intraindividual representational system and as an interindividual communication system. Clearly, both aspects are interrelated, but we can separately consider the role they play in developing the child's understanding of other minds.

INTERDEPENDENCE OF LANGUAGE AND THEORY OF MIND IN DEVELOPMENT

The development of children's understanding of other minds, or their theory of mind, is inherently intertwined with language development. Admittedly, some theorists ascribe a fairly limited role to language. For example, those with an innate modular view of cognitive skills (Baron-Cohen, 1995) claim that theory of mind is inborn but children cannot express the skill until their language is sufficiently mature. Other researchers argue that the understanding of other minds depends on domain-general cognitive operations (Frye, Zelazo, & Palfai, 1995) that require language for their implementation. Yet others consider language to be an inherent aspect of children's social interaction, providing children with information necessary for the construction of theory of mind without playing any fundamentally causal role (Gopnik & Wellman, 1994).

Notwithstanding these views, there are many researchers and theoreticians who ascribe a more fundamental role to language as a route into other minds, especially during the toddler and preschool period. In line with this position, our aim is to explain what it is about language that influences theory-of-mind development. Language and theory of mind are intricately related, and their development is intertwined during the first 5 years of life. Both language and theory of mind develop along their own developmental paths, yet influence one another (Malle, 2002). As described above, both are multifaceted systems, each composed of a number of components. Their interdependent relation is thus complex because there are possibilities for different relations among the different components, for changes and modifications in these relations over developmental time, and for variations in relations in different individuals.

The interdependence of the two systems is evident in that both language and theory of mind are systems used for representation and communication. Language enables us to communicate our own representations of reality, or points of view, and to hear about others' points of view. Theory of mind involves an understanding that people have varying points of view, that is, different representations of the world, that underlie their communicative exchanges.

Both language and theory of mind are central to human social interaction. They are acquired during the preschool years and both are well developed by about 5 years of age. This is not to say that the development ceases there. On the contrary, having acquired the fundamental skills in both systems, children are equipped with the rudimentary tools that help them to become competent participants in social life. These skills represent the means to interact with others in increasingly sophisticated ways and to interpret the daily discourse, nonliteral language, literature, and perspectives expressed in different art forms.

HOW LANGUAGE PROVIDES A ROUTE INTO OTHER MINDS

First we consider language as an interindividual communication system and investigate the role of social discourse in opening up a route into other minds, particularly during the toddler and preschool period. Then we consider language as an intraindividual representational system and investigate how children's own developing linguistic ability in semantics and syntax further develops their understanding of mind during this period. Finally, we consider pragmatics, where the interrelation of language and theory of mind is most evident and where it is clear that the relationship is bidirectional.

Before the preschool period, infants' social awareness (i.e., their attunement to people as a special category of objects) drives language acquisition. During the preschool years, children's mastery and use of mental verbs helps develop their understanding of other minds. Beyond the preschool years, children's metarepresentational ability (i.e., the ability to reflect on others' mental states) facilitates their comprehension and production of more complex language in discourse. Thus, language and theory of mind develop together, with advances in one sphere facilitating and predicating advances in the other.

Social Discourse Context

The kind of language children are exposed to as participants and observers in communicative exchanges influences their understanding of other minds. It is obvious that the linguistic environment will affect children's own linguistic abilities, and, conversely, their linguistic ability may affect their environment in terms of the kinds of communications they receive.

Although much about others' minds can be inferred from behavior, gestures, and facial and emotional expressions, it is primarily through linguistic interaction that we learn about others' feelings, beliefs, desires, and intentions. In talking with parents, siblings, and friends and in listening to conversations between others, children learn about others' minds. Both theory of mind and language develop as children actively participate in the social world, in which participation is contingent upon communication (Nelson, 2005).

Dunn's research team was first to demonstrate the effect of the linguistic environment on young children's understanding of other minds (Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991). They showed that maternal conversation about people's feelings and about causal relations when children were 33 months old predicted successful false-belief understanding at 40 months of age. Moreover, the quality of the relationship between communicative participants has been shown to have an effect on the efficacy of conversation in developing understanding of the mind (Dunn & Brophy, 2005). Ruffman, Slade, and Crowe's (2002) findings from a longitudinal study of children (mean age 36 months at the first time-point) further substantiate the claim that the quality of maternal discourse determines the child's theory-of-mind development. In their study, the use of mental-state terms (that is, terms expressing desires, beliefs, emotions, etc.) by mothers predicted children's theory-of-mind skills 1 year later, even when controlling for children's earlier theory-of-mind abilities, language skills, and own use of mental-state terms.

However, Harris (2005) points out that mothers' use of mental-state terms might be just a readily detectable aspect of a style of talk in which the mother introduces diverse points of view into the conversation, and that it is the latter that is the critical element in engendering understanding of other minds. Evidence in support of this argument comes from training studies, where one can experimentally manipulate the linguistic input children receive. Peskin and Astington (2004) showed that stories involving tricking and hiding, containing many mental-state terms, are no more effective in children's developing false-belief understanding than the same stories without mental-state terms. In fact, stories lacking explicit mental-state terms were more effective in training children to explain behavior based on their identification of false beliefs. In further support of Harris's argument, findings from Lohmann and Tomasello's (2003) training study demonstrate that an emphasis on different points of view, without using mental-state terms, provides a sufficient input to generate an improvement in children's performance on theory-of-mind measures.

Thus, it is through conversational interactions that children become aware that other people may think, know, like, and want different things from what they themselves do. The contribution of the linguistic environment, the interindividual nature of linguistic interaction, is irrefutable. Yet, its role as a route into other minds is only a part of a larger developmental picture. In order to see the full picture, one also needs to consider the intraindividual aspects of language and the child's own developing semantic, syntactic, and pragmatic competence.

Semantics

Semantic knowledge is knowledge of meaning expressed through language. Studies have shown strong correlations between children's performance on false-belief tasks and measures of semantic skills, such as receptive vocabulary. For instance, using data pooled from a number of studies, Happé (1995) demonstrated that 3- to 4-year-olds' verbal ability measured by the British Picture Vocabulary Scale (BPVS; Dunn, Dunn, Whetton, & Pintilie, 1982) was highly correlated with their false-belief task performance and made a significant contribution to variability in false-belief scores independent of age. Furthermore, Cutting and Dunn (1999) showed that children's scores on the same measure correlated with aggregate performance on a range of false-belief tasks and with a test of emotional understanding, even when age and family background were accounted for. More striking evidence comes from a recent meta-analysis of the relation between language and false-belief understanding (Milligan & Astington, 2005) which shows that in studies involving almost 600 children approximately 25% of the variance in false-belief task performance is accounted for by children's semantic knowledge.

Children's high scores on general semantics measures may be an indication of their participation in rich conversational exchanges (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991), thus providing further evidence of the role of the linguistic environment in theory-of-mind development. When focusing on the intraindividual aspects of language, however, a particular interest falls onto the child's semantic knowledge of specific lexical items—that is, terms used to refer to mental states.

The mental world is a world of unobservable abstract entities, such as beliefs, desires, intentions, and emotions, that are variously revealed in facial expressions, talk, and behavior. Language provides a means of abstracting the underlying concepts from the ongoing stream of social interaction, as children acquire the mental-state concepts that are semantically encoded in the language of their culture. When they are 2 and 3 years old, children acquire mental-state terms: first perception, emotion, and desire terms (e.g., *see, look, happy, sad, like, love, want*), and then cognition terms (e.g., *know, think, remember*) (Bartsch & Wellman, 1995; Bretherton, 1991). Although some researchers argue that production of these terms is itself evidence of understanding other minds, others look for relations between production and/or comprehension of these terms with performance on experimental tasks, such as false-belief tasks. For example, Moore, Pure, and Furrow (1990) showed that there is a relation between children's comprehension of terms such as *think* and *know* in an object-finding task and their performance on tests of false-belief understanding. Production of such mental-state terms in naturalistic play is also associated with false-belief task performance (Hughes & Dunn, 1998).

A critical question is: How do children acquire the concepts from language? Although children acquire mental terms from their participation in conversation, they are not passive recipients of mentalistic concepts but, rather, actively involved in their construction. It has been argued that language allows for a level of abstraction that can support abstract mental-state concepts, perhaps via a process involving analogical reasoning and inference (Baldwin & Saylor, 2005). Others argue that young children may use the mental-state terms that they hear other people using—without at first having any conceptual understanding of mental states. This understanding is gradually developed from using mental-state terms in familiar contexts and comparing the ways oneself and others use a term (Montgomery, 2005; Nelson, 2005).

However this debate is resolved, it is clear that children's ability to talk about unobservable mental states is an important stage in their developing ability to think about other minds and to interpret others' behavior by reasoning about mental states. This reasoning requires metarepresentation, that is, the representation of mental attitudes toward mental contents. As language develops, increased resources in syntactic structures provide the format required for such representation, and thus, it is argued, syntactic development facilitates mental-state reasoning (J. G. de Villiers, 2005).

Syntax

Syntactic knowledge involves knowledge of appropriate linguistic structure. In the previously mentioned meta-analysis of the relation between language and false-belief understanding (Milligan & Astington, 2005) we also examined the correlation of syntactic measures and false-belief tasks. In a different subset of studies, again involving almost 600 children, 28% of the variance in false-belief task performance was accounted for by children's syntactic knowledge. This accords with our earlier finding of a strong and predictive relation between syntactic ability and false-belief understanding in a longitudinal study (Astington & Jenkins, 1999).

As in the case of semantics, however, when considering the relation between theory of mind and syntax, a particular focus falls on mental verbs. More precisely, the focus is on the specific syntactic structures that provide the format required to represent false beliefs, that is, sentential complements. Mental verbs occur as the main verb in a complex sentence that contains an embedded clause—the sentential complement. For example, "Maxi thinks *the chocolate is in the cupboard*" (sentential complement in italics). Importantly, the whole complex sentence, including main and embedded clauses, can be true even if the embedded clause expresses a proposition that is false. Mastering such syntactic structures is essential for the expression of false beliefs.

Children use complement constructions almost as soon as they start to produce mental verbs—that is, at 2 years of age (Bartsch & Wellman, 1995; Bloom, Rispoli, Gartner, & Hafitz, 1989). Only a small number of verbs are widely used with complements at first, primarily *see, look, think,* and *know.* Importantly, Diessel and Tomasello (2001) show that this early use is formulaic and argue that it does not provide evidence of mastery of complement syntax. The comprehension of complements, assessed by children's memory for complements, is not mastered until a year or two after the first use, when it correlates with children's performance on false-belief tasks. Furthermore, the comprehension of such complements predicts this performance in a longitudinal study (J. G. de Villiers & Pyers, 2002). Mastery of complement syntax supports metarepresentation by providing the format for the representation of mental attitudes toward mental contents.

Findings from children with autism and deaf children provide even more striking support for the importance of mental verbs as a route into other minds. Even though children with autism have deficits in theory of mind and mental verb use, some of these children do eventually develop false-belief understanding, and their mastery of sentential complements is the best predictor of success (Tager-Flusberg & Joseph, 2005). Likewise, non-native deaf signers are delayed in theory-of-mind development relative to deaf children who acquire sign as a native language. This may be because the latter group has early access to mental verbs via sign language that the former group lacks. Once again, as with children with autism, mastery of sentential complements is the best predictor of deaf children's success on false-belief tasks (P. de Villiers, 2005).

We have demonstrated that children's mastery of specific aspects of semantics and syntax is predictive of (or supports) their successful performance on theory-of-mind tasks. Yet, this acquisition takes place in a lively context of language in use, and, indeed, the dynamics of the mutual influence of language and theory-of-mind development are nowhere as apparent as in the domain of pragmatics.

Pragmatics

Pragmatic competence underlies the ability to use and interpret language appropriately in communication with others, which requires some awareness of the mental states of one's communicative partners. Thus, linguistic (i.e., pragmatic) competence and theory of mind are necessarily related, and there is evidence of their close association from infancy onward. Infant intersubjective social behaviors, such as imitation, gaze following, joint attention, and so on, are sometimes referred to as "early" or "implicit" theory of mind (e.g., Bretherton, 1991). In a longitudinal study of 9- to 15-month-old infants, Carpenter, Nagell, and Tomasello (1998) found high correlations of such behaviors with measures of the infants' communicative competence, for example, in producing words and gestures and comprehending language. By 18 months of age, infants can use the direction of a speaker's gaze to infer the referent of a novel word (Baldwin, 1993). Baldwin demonstrated that, upon hearing a new word in the presence of two unfamiliar objects, infants associate the word with the object at which the speaker is looking. They do so even in a condition in which their attention is directed to the other object. That is, they take the speaker's gaze direction as an indication of his or her intent to refer. It is worth noting that children with autism, who appear to lack these fundamental intersubjective abilities, cannot learn new words in this way (Baron-Cohen, Baldwin, & Crowe, 1997). For typically developing children, however, the precursors of theory of mind facilitate language development. From this point on, language develops rapidly and becomes the route into a more explicit understanding of other minds.

Later in the preschool years, children's pragmatic competence is related to experimental measures of mental-state reasoning, such as false-belief task performance. For example, Dunn and Cutting (1999) showed that false-belief understanding was related to connected and successful communication between friends in a naturalistic play situation. Beyond the preschool years, children's pragmatic ability to comprehend and use figurative language and their skills at constructing narratives are related to further developments in theory of mind. For example, irony comprehension is related to higher-order theory-of-mind skills (Filippova, 2005; Happé, 1993). Furthermore, Comay (2005) shows that advanced theory-of-mind skills play a unique role in narrative development beyond a common link with language because the successful narrator has to represent the story characters' mental perspectives within the story frame as well as maintain an awareness of the listener's perspective outside of the story frame.

CONCLUSION: THE BEAUTIFUL LIE

In conclusion, innate abilities underlying social perception provide a base for language to develop. Language is first used in social interaction and is then internalized as a representational device. Participation in conversation leads to awareness of mental states, while children's own syntactic and semantic abilities facilitate metarepresentational interpretations of human behavior. Metarepresentational ability allows children to represent false propositions—beautiful lies. Why "beautiful"?—it may seem odd to describe a disapproved activity with such an approving term. However, even if lies themselves are not to be celebrated, the metarepresentational ability that is expressed in many ways (including lie-telling) is an enormously significant development. As Pugh (2002, p. 7) put it: *"This is how to make songs, create men, paint pictures, tell a story."* These abilities are the essence of humanity. We have good reason to celebrate the development that underlies them.

ACKNOWLEDGMENTS

We are very grateful to Sheenagh Pugh for permission to reprint her poem "The Beautiful Lie." We are also grateful to the Natural Sciences and Engineering Research Council of Canada for financial support of our research.

REFERENCES

Astington, J. W., & Baird, J. A. (2005). Introduction: Why language matters. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 3–25). New York: Oxford University Press.

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory of mind development. *Developmental Psychology, 35*, 1311–1320.

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20*, 395–418.

Baldwin, D. A., & Saylor, M. M. (2005). Language promotes structural alignment in the acquisition of mentalistic concepts. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 123–143). New York: Oxford University Press.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind.* Cambridge, MA: Bradford Books/MIT Press.

Baron-Cohen, S., Baldwin, D. A., & Crowson, M. (1997). Do children with autism use the speaker's direction of gaze strategy to crack the code of language? *Child Development, 68*, 48–57.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York: Oxford University Press.

Bloom, L., Rispoli, M., Gartner, B., & Hafitz, J. (1989). Acquisition of complementation. *Journal of Child Language, 16*, 101–120.

Bretherton, I. (1991). Intentional communication and the development of an understanding of mind. In D. Frye & C. Moore (Eds.), *Children's theories of mind* (pp. 49–75). Hillsdale, NJ: Erlbaum.

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63*(4, Serial No. 255).

Comay, J. (2005). *Individual differences in narrative perspective-taking and theory of mind: A developmental study.* Doctoral dissertation in preparation, University of Toronto, Canada.

Cutting, A. L., & Dunn, J. (1999). Theory of mind, emotion understanding, language and family background: Individual differences and interrelations. *Child Development, 70*, 853–865.

de Villiers, J. G. (2005). Can language acquisition give children a point of view? In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 186–219). New York: Oxford University Press.

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief understanding. *Cognitive Development, 17*, 1037–1060.

de Villiers, P. (2005). The role of language in theory-of-mind development: What deaf children tell us. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 266–297). New York: Oxford University Press.

Diessel, H., & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics, 12*, 97–141.

Dunn, J., & Brophy, M. (2005). Communication, relationships, and individual differences in children's understanding of mind. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 50–69). New York: Oxford University Press.

Dunn, J., Brown, J., Slomkowski, C., Tesla, C., & Youngblade, L. (1991). Young children's understanding of other people's feelings and beliefs: Individual differences and their antecedents. *Child Development, 62*, 1352–1366.

Dunn, J., & Cutting, A. L. (1999). Understanding others, and individual differences in friendship interactions in young children. *Social Development, 8,* 201–219.

Dunn, L. M., Dunn, L. M., Whetton, C., & Pintilie, D. (1982). *British Picture Vocabulary Scale*. Windsor, UK: NFER-Nelson.

Filippova, E. (2005). *Development of advanced social reasoning: Contribution of theory of mind and language to irony understanding*. Unpublished doctoral dissertation, University of Toronto.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development, 10*(4), 483–528.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press.

Happé, F. G. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition, 48,* 101–119.

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*, 843–855.

Harris, P. L. (2005). Conversation, pretense, and theory of mind. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 70–83). New York: Oxford University Press.

Hughes, C., & Dunn, J. (1998). Understanding mind and emotion: Longitudinal associations with mental-state talk between young friends. *Developmental Psychology, 34*, 1026–1037.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27,* 236–248.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false-belief understanding: A training study. *Child Development, 74*, 1130–1144.

Malle, B. F. (2002). The relation between language and theory of mind in development and evolution. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 265–284). Amsterdam: Benjamins.

Milligan, K. V., & Astington, J. W. (2005). *The relation between language and false belief understanding: A meta-analysis*. Manuscript submitted for publication.

Montgomery, D. E. (2005). The developmental origins of meaning for mental terms. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 106–122). New York: Oxford University Press.

Moore, C., Pure, K., & Furrow, D. (1990). Children's understanding of the modal expressions of speaker certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development, 61*, 722–730.

Nelson, K. (2005). Language pathways into the community of minds. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 26–49). New York: Oxford University Press.

Peskin, J., & Astington, J. W. (2004). The effects of adding metacognitive language to story texts. *Cognitive Development, 19*, 253–273.

Pugh, S. (2002). The beautiful lie. In *The beautiful lie* (p. 7). Bridgend, Wales: Seren.

Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Development, 73*, 734–751.

Tager-Flusberg, H., & Joseph, R. M. (2005). How language facilitates the acquisition of false-belief understanding in children with autism. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 298–318). New York: Oxford University Press.

14

Representation of the Interlocutor's Mind during Conversation

MARJORIE BARKER
T. GIVÓN

BACKGROUND

During natural face-to-face communication, people engage simultaneously in a great number of social, communicative, and cognitive tasks. Among these, the two most prominent ones are the interaction between the interlocutors and the management of the information flow. In other words, people engaged in conversation are simultaneously tracking both the social collaboration that allows their interchange of ideas and the gradual development of a body of information as it flows back and forth between them. It seems obvious that the two tasks cannot be totally divorced from each other, since the interpersonal dynamics serve as the matrix within which the information flow takes place. This matrix involves not only the situational aspects of the interaction—speaker and hearer, time and place—but also the current state of knowledge and intention in the mind of each interlocutor at any given moment (Givón, 2001b, 2005; Grice, 1968/1975).

In both these aspects of human communication, grammar is used systematically as a context-cuing, perspective-shifting device (Givón, 2001b, 2002, 2005). Grammar indexes the relative importance of conversational referents, places them in space and time, and conveys the speaker's opinion about the reliability and desirability of the information being conveyed, all of which change continually as conversation proceeds. As speakers use grammar to communicate, they rely, constantly and systematically, on their own constantly shifting mental models of the hearer's belief and intention states. At the same time, they manipulate the hearer's construction of *their* constantly shifting mental representations.

There is ample evidence in natural language data that tracking the listener's physical and social position, as well as the listener's perspective on information being communicated, is a prime concern of speakers; entire portions of the grammar of any language are devoted to these things (Givón, 2001a). The features that demonstrate this concern in language are ubiquitous but subtle. One such field is *deixis*—referential (*I, you, this, that*), spatial (*here, there*), temporal (*now, then, today, yesterday*), and social (*Mr. Smith* vs. *John* vs. *honey*) identification of the current speech situation. Such reference shifts constantly with the shift of conversational turns and speech situation: *I* to one speaker does not have the same referent as *I* to the other speaker; *this* and *that* may identify the same referent, depending on the perspective taken.

The existence of mental models of the interlocutor's knowledge is also the only explanation for the linguistic phenomenon of definiteness: the choice between "a woman" and "the woman" is based on the speaker's perception of a difference in the *listener's* mind. If I think my listener will be able to identify the person I am talking about, I use a definite article, "the woman"; otherwise, I bring *the same referent* into the conversation as "a woman." To argue that, for example, speakers use "a" when first introducing a word and "the" subsequently, at no point considering the hearer's familiarity with the term, does not explain why some referents are introduced initially as known.

An equally conspicuous example of implicit mental models of the interlocutor's epistemic and intentional states involves the grammar of speech acts in ordinary conversation (Givón, 2001b). To use an interrogative appropriately, for example, the speaker must believe that the hearer might know the answer and be willing to share the answer, or at least speculate about it. The appropriate use of a declarative involves the speaker's belief that the hearer may not be aware of what the speaker is about to say and would welcome knowing it. Indeed, to initiate conversation at all, speakers make judgments about the other person's willingness and ability to participate in the desired interaction. These mental models about the hearer's presumed belief and intention states must shift constantly as conversation proceeds, as new information is exchanged. And the most straightforward way to account for the felicitous use of the grammar of speech acts during conversation is to assume the existence of such mental models.

Constantly shifting, and largely implicit, mental models of the interlocutor's current belief and intention states are characteristic of both conversational and narrative discourse. Furthermore, the speaker's shifting perspective about the hearer's mental representation does not depend on shifting conversational turns. The status of referents within a single turn changes as the speaker continually revises his or her estimation of what the hearer knows. As a simple illustration, consider one speaker's use of the grammar of referent tracking over the course of one conversational turn:

> VICKI: OK. And after that she made *a fire* . . . umm she picked up *a little, bad old tin pail*, and she went over next to *the lean-to*, the white bucket of *water*, and she poured *the water* into *the pail* and then took *the pail* back and put it on *the fire*.

The use of the indefinite noun phrases "a fire," "a . . . pail," and "water" presupposes that the hearer has no prior episodic representation of these three referents. But the subsequent use of "the water," "the pail," and "the fire" assumes that the hearer now has such mental representations.

The very act of speaking assumes an ability on the part of the hearer to develop a mental model of the speaker's perspective—and this assumption is, in turn, a theory about the listener's mind. The grammatical features described above, and others like them, serve as linguistic evidence that modeling the interlocutor's mind is important to speakers and hearers. The prevalence of these indicators suggests that it should be possible to elicit empirically some acknowledgment by conversational participants of the mental modeling taking place during conversation.

Though the two main aspects of communication—interaction and information—cannot be totally separated, their study has generally been split into two methodological paradigms. The study of the information flow of communication has focused almost exclusively on either the produced text (in linguistics; e.g., Chafe, 1994; DuBois, 1987; Givón, 1994; Tomlin, 1991) or on text comprehension (in psychology; e.g., Anderson, Garrod, & Sanford, 1983; Ericsson & Kintsch, 1995; Gernsbacher, 1990; Trabasso, Suh, & Payton, 1995). The study of conversational interaction, on the other hand, began with a focus on the social aspect of the interaction, primarily the turn-taking system (e.g., Goodwin, 1988;

Sachs, Schegloff, & Jefferson, 1974). This methodological cleavage has persisted in spite of numerous studies in both the discourse and acquisitional literature that suggest a strong interaction between the interactional and information-flow aspects of communication (e.g., Chafe, 1997; Coates, 1997; Ervin-Tripp & Kuntay, 1997; Goodwin, 1995; Wilkes-Gibbs & Clark, 1992).

An absolute separation between the social and informational study of natural communication is theoretically indefensible and methodologically unwise. The information flow and the speech situation are inextricably bound together. What is more, both are equally "cognitive"; what is relevant in both is not some "objective" reality but rather the conversational participants' mental representation of whatever that reality may be. Any objective reality, be it situational or textual, that does not attain some type of mental representation is irrelevant to the process of communication. And the very same mental representation systems must be involved in the processing of both aspects of communication, since the task of language processing is at heart a matter of interpreting the mind and intention of the interlocutor.

The experiment described here was an attempt to tease out people's mental representation of their interlocutor's perspective during conversational interaction. In this way, the study probes the representation in episodic memory of the interrelated factors we suspect to be equally salient during conversation: both the communicated contents and the communicative interaction. In order to keep the informational content of the conversations fairly constant, we asked participants to discuss a video they both had just finished watching. We predicted that asking participants to find differences in what they had just viewed would cause speakers to pay closer attention to both the information provided by their partner and to their partner's mental attitude toward this information.

METHODOLOGY

Material and Participants

Participants (*N* = 14 undergraduate students) watched a 6.25-minute videotape, "The Chicken Story," developed and used for two other experimental projects (Dickinson & Givón, 1997; Givón, 1991). The film shows a man and a woman, working separately and interacting, in a rural setting. The video was filmed in Eugene, Oregon, and the actors speak Swahili. There are no subtitles, allowing viewers to develop their own interpretations of the content of the actors' interaction. Fifty-eight states/events in the video were established by the earlier experiments as baseline clauses for recall, because they were mentioned by at least 7 of 10 participants who described the action aloud while watching it.

Participants watched the video at the same time, but separated from one another, on individual monitors. Prior to viewing, they were told to pay close attention and memorize the story. After watching the film, they were brought together and told (incorrectly) that the films they had just seen were similar in broad outline but not completely identical. They were then asked to talk with each other about the film in order to find out as much as they could about the video *the other person* had seen. These conversations about the film were recorded. Participants were then debriefed individually and asked to recall in as much detail as possible the conversation they had just had. These recall monologues were also recorded.

Analysis of Transcripts

The experiment produced two sets of transcripts, one set of seven dyadic conversations and the other of 14 individual recalls of the conversations.

The conversations were not the prime target of this investigation; our interest was in the participants' mental representation of the interaction: could they report on the nature of the interaction as well as the information involved? For this reason, the focus of the analyses was to compare the recall transcripts to the original conversations and assess participants' accuracy in their attribution of who said what, how confident they were in saying it, and differences of opinion—in other words, their accuracy in understanding the other person's perspective. For this reason, the recall transcripts were more extensively analyzed than the conversation transcripts.

Conversation transcripts were analyzed to assess the number of baseline events mentioned by each pair during the course of conversation and to determine who mentioned each event, data that served as criteria for the recall analysis.

The text of individual recall transcripts was divided into (1) recall of the contents of the video and (2) recall referring to the dynamics of the conversation itself. The following brief passage from a representative recall transcript illustrates this separation. Portions of the text referring to the conversational interaction are **boldfaced** and those referring to the video are in plain type.

> **. . . the conversation I had with Vicky. First of all she started out by saying, saying** what the man was wearing, **and that** he was carrying three farming utensils plus a hatchet, **and, actually, no, she started out asking me what** he was wearing. **And I said that** he was wearing red shorts and a

> white T-shirt and no thongs or anything. **And she said that** he was wearing red shorts and a white T-shirt, but that he had flip-flops on.

The proportion of clauses in each recall transcript referring to video contents versus conversational interaction was established. Recall transcripts were also analyzed to determine the number of video baseline events mentioned in the subjects' recollection of the conversation. In addition, these transcripts were compared, for each pair, to the conversation they were attempting to recall. We computed accuracy of recall for (1) which speaker produced a particular clause in the original conversation and (2) the speech-act modality used, divided broadly into "certain" versus "uncertain."

Uncertain modality found in the conversation transcripts was marked in five different ways: as interrogative speech acts (*what was the guy wearing?*); nonfactive or negative cognition verbs (I *think* mine had two; I *don't remember* seeing her fill it); nonfactive or negative perception verbs (it *looked like* cheese; I *didn't see* her put anything in the water); nonspecific reference (*whatever* she had wrapped up, in the towel, she put that down; and brings back some, um, *something*); and uncertain quantifiers or adverbs (there're *maybe* seven or eight of them; she *kinda* looks disgusted).

RESULTS

Conversations

During conversation, subjects collaboratively mentioned on average 37 (64%) of the 58 baseline events of the video. Averaging over all pairs, each partner produced about a third of the baseline clauses mentioned in conversation, with the remaining third spoken by both partners. There is wide variation, though, in the proportion of baseline material handled by any individual speaker, ranging from 13% to 55% (SD = 13.2%). Interestingly, and perhaps predictably, the first speaker of each pair tends to handle significantly more of the informational content of conversation (M = 39%) than the second speaker (M = 27%); t = 3.75, df = 12, $p < .01$. The difference between male and female speakers was not significant.

Recall

In recalling the conversations they just participated in, subjects produced speech that referred to the collaborative aspect of their discussion as well as speech recounting the information they had exchanged.

On average, about one-fourth of recall clauses referred to the circumstances of interaction. The three most common features of interaction mentioned in the recall transcripts were identification of the speaker (*she* said, *I* said), epistemic/modal qualification of the recalled information ("*I think* that she said," "and *maybe*," "*we couldn't remember*"), and identification of the speech act or modality used in the conversation ("she started by *asking*," "we kind of *disagreed*," "I thought I *saw*"). We explored the extent and accuracy of each of these features.

Interestingly, participants had no apparent trouble attributing every chunk of remembered content of their conversation to either themselves or their interlocutor. Our initial assumption was that this attribution need not be accurate, because the two people saw the same video and agreed on most of the details of what they had seen. Their conversation was thus extremely collaborative, and it seemed there would be no point in mentally representing two separate points of view. Nonetheless, speakers were 86% accurate overall in remembering who said what in the original conversation. (Only two speakers had a significant amount of the hedged attribution *"we"*; these instances were counted as neither accurate nor inaccurate.) This result indicates that speakers have an interest in distinguishing speaker identity irrespective of agreement about the conversational content.

Speakers were also surprisingly certain in recalling the speech acts that occurred in the course of their conversation. Epistemic and modal markers of uncertainty in connection with conversational recall were rare. There were 557 instances of quotative verbs in the recall transcripts ("I said," "she agreed," "he asked"); of these, only 56 (10%) were explicitly qualified by some lower-certainty marking ("I *think* he said, "I *don't remember if* we talked about that," "she *might have* asked"). Thus, regardless of their actual memory accuracy, subjects were—at least if one judges by their overt verbal use of modality—overall quite confident in their recall.

It is more difficult to assess the accuracy of subjects' recall of the specific speech-act modality used in their original conversation, for several reasons. First, there are virtually no quotative speech-act verbs in the conversations themselves. People simply do not say things like "I'm asking you what he's wearing" or "I say there was some kind of noise in the background." Rather, they say, "What was he wearing?" and "There was some kind of noise in the background." Also, multiple quotative expressions in the recall texts—*say, talk, mention, describe, tell, remember, comment, recall*—may be used to refer to the very same declarative speech act of saying. And while the predominant quotative verb in the recall transcripts is "say" (281/557 = 50%), there is no principled way of ruling "say" as the "correct" quotative usage or preferring it to any of the other verbs used.

It is obviously problematic to determine the accuracy of conversationalists' recall for speech-act modality if we need to make judgments about whether the original speech act should be considered *saying* rather than *talking, commenting,* or any of the other expressions used by speakers to recall conversational speech. One possible line of analysis would be to lump together all the declarative speech acts in the conversations, contrast them with the interrogative ones, and then compare all the uses of "ask" in the recall transcripts to see if they match the use of interrogatives in the respective conversations. But to separate out only interrogatives would reduce our sample to less than 5% of the total use of quotative verbs (27 instances of questions out of 557 total recalled speech acts).

As an alternative, we elected to assess a somewhat rougher speech-act variable, that of the *epistemic modality* of speech acts. We divided epistemic modality of conversational speech acts into certain (*realis*) versus uncertain (*irrealis*), taking certainty to be the default case (i.e., if not specifically marked by uncertain grammar, as described in the "Analysis of Transcripts" section, utterances were considered to be expressed as certain).

There are two kinds of errors: (1) an instance of uncertain modality in conversation that was later recalled as certain and (2) an instance of certain modality that was recalled as uncertain. There were only five exemplars of the latter in the recall transcripts, three produced by a single speaker and the other two by two participants recalling the same utterance. All five concerned interchanges where there was a difference of opinion in the original conversation as to the video content. An example is the following exchange, where G's original statement in the conversation was certain, but later recalled as uncertain by both participants:

Conversation:

G: Since she built the fire?

D: Yeah . . .

G: *Then she put it out*, for some reason . . .

D: Oh, no, she put, mine, with mine she put, uh . . . she put water on there and had it boil.

G: Oh, ah . . .

D: Boiling . . .

G: That's what she's doing.

Recall, separately by both participants:

G: *I thought she put the fire out*, but apparently she was boiling water to cook chicken.

D: *He thought that it was put out*, and I thought that it was still going.

The other three exemplars of certain utterances later recalled as uncertain were all produced by one participant, who differed with her partner's interpretation of the video in each instance in the original conversation:

Conversation:	*Recall*:
L: It was a bag.	V: She thought it was a bag.
L: She made two cuts.	V: She made two cuts, she thought.
L: She was bringing him lunch.	V: She thought she was bringing him lunch.

In each of the cases above, the recall that the original speaker "thought" something happened is used to highlight a contrast, either between the two conversational partners' memory of the video or between the speaker's original interpretation and a change in that interpretation during the course of interaction. In each case, there is an identification of what the person said with what he or she thought—of speech as evidence of mental representation.

We limited further analysis to instances where the speakers expressed uncertainty. If listeners are tracking their partner's mental state during the course of conversation, they should be paying particular attention to explicit expressions of uncertainty and might be expected to recall these as less confident.

We matched the original conversation with transcripts of its recall by both speakers involved and determined how accurately each instance of uncertainty was recalled. Instances of uncertainty in the conversation that were not mentioned by either partner in recall were excluded from analysis. The remaining instances were grouped in four categories, as shown in the examples below.

1. Correctly recalled as uncertain by the original speaker.

   Conversation:

   M: and then a woman came, tall and uh . . . white blouse and pink . . . skirt.

   D: Yeah.

M: And . . . came and . . . *seems to be* yelling at him about something.

Recall:

M: I had mentioned that she *seemed to be* yelling.

2. Correctly recalled as uncertain by the original listener.

Conversation:

L: She was trying . . . trying to, y'know, slit the chicken's throat . . .

V: Yes, she was, the chicken was just a wild little guy . . .

L: Yeah, *I don't think* she really had the heart to do it.

Recall:

V: And Lori commented that *she didn't really think* that the woman would do it.

3. Incorrectly recalled as certain by the speaker.

Conversation:

L: OK, of course she . . . she walked over to where the fire pit was.

V: Uh huh.

L: She put wood into the fire pit . . . and then she uh . . . she started the uh . . . fire . . . and she put . . . *I thought* she put a little bit of kindling in to get it going?

Recall:

L: Then I said that she . . . she took the wood. I said went over to the fire and started it, and went to get, she, she lit it and *then she put* a little more brush on it to get it going.

4. Incorrectly recalled as certain by the listener.

Conversation:

L: OK.

T: Uh . . . sh . . . she seems to use a . . . she *seems to* have a little . . . uh big butcher knife.

L: Yeah.

Recall:

L: He said something about the knife, she *had* this big butcher knife is what he said.

The modality of uncertain speech acts in the conversation was recalled correctly by the original speaker 94% of the time (73 out of 78 instances) and by the listener 85% of the time (64 out of 75 instances). These results confirm, albeit in a limited way, that both participants are accurately tracking the speech-act modality of conversation.

DISCUSSION

We have shown that subjects recalled two features of their conversation with a high degree of accuracy—attribution of quoted information to the correct speaker (86%) and epistemic modality of the speech act (90%). The first is somewhat surprising because the conversations we provoked were extremely cooperative, and speakers had little disagreement about the contents of the video. It might be expected, therefore, that it would hardly matter who said what, and thus this information would not be easily remembered. The second finding is surprising, too, because the distinction between certain versus uncertain modality is often subtle; and, again, it is not easy to see why it would be important in the present experimental context. That both aspects were accurately recalled is evidence that both the informational content and the interactional aspect of face-to-face communication receive consistent conscious episodic representation.

Furthermore, we have shown that the mental representation of conversation is integrated—memory for the contents and memory for the interaction itself are closely interlaced. These findings in themselves are not too surprising and may have been predicted from either general consideration or anecdotal evidence.

What is more interesting is the striking contrast between the type of interactional information that consistently achieves conscious representation and the type that just as consistently does not. A whole range of interactional information, described earlier as *necessarily* activated—readily, copiously, and perhaps automatically—during face-to-face conversation, never achieves the conscious representation necessary for verbalization and is entirely absent from the recall transcripts in this experiment. This is the information associated with the use of grammatical constructions and morphemes whose felicitous deployment is indispensable to human communication as we know it, be it face-to-face or narrative—information like deixis, definiteness, and speech-act grammar. Here, again, the explicit mention of this kind of information simply does not occur in ordinary conversation—people don't report their reasons for using *the* rather than *a* ("I told him about *a* woman in the video, not *the* woman, because I wasn't sure whether he had seen her") or their motivation for initiating specific interactions ("I notice you don't seem to be doing anything right now, and you made eye contact with me, and I'm assuming, since we are both in the United States, that you understand English, so I'm going to try to initiate an interaction. . . .") Nonetheless, the consistent use in conversation of these grammatical features argues that the mental modeling in spoken interaction responsible for their application is continually taking place. The fact that it is never verbally reported by our participants in the transcripts of conversational recall suggests that it takes place automatically and unconsciously.

Here is a tentative explanation for why the vast reservoir of mental models that speakers construct about their interlocutor's mind is implicit rather than conscious, verbal, and explicit. First, grammar is an automated speech processing device, sensitive to extremely local communicative contexts (Blumstein & Milberg, 1983; Givón, 1979, 1989, 1992, 1994; Kintsch, 1992; Neville, Mills, & Lawson, 1992). Second, the relevant mental states—of both speaker and hearer—shift constantly from one clause or even one word to the next. Third, preserving a longer-term—global—conscious episodic representation of such context-specific mental states is adaptively useless, since their relevance decays rapidly. And finally, preserving a long-term mental trace of these transitory, local mental states may indeed be adaptively damaging, since they would then interfere with the more relevant mental models of the hearer's newer, current mental state.

The dichotomy between conscious (attended) and implicit (automated) information processing, and thus mental representation, has been a major theme in cognitive psychology and neuroscience for over three decades (e.g., Kahneman & Treisman, 1984; Posner & Snyder, 1974; Schneider & Shiffrin, 1977). This dichotomy has been demonstrated extensively in visual information processing (DeSchepper & Treisman, 1996; Fernández-Duque, 1999; Posner & Peterson, 1990; Treisman & DeSchepper, 1996), but also in the processing of written words (Sieroff & Posner, 1988) and in the recall of declarative verbal information (Nissen & Bulleme, 1986). What is more, attentional activation—working memory—can itself be either conscious or implicit (Gathercole & Baddeley, 1993; Shallice, 1988).

This interpretation of our results may seem to stand in contrast to the claims of Keysar and others that the perspective of the interlocutor is not initially considered in formulating speech, and only later conscious revision permits speakers to tailor their utterances to a particular listener (e.g., Barr & Keysar, Chapter 17, this volume; Horton & Keysar, 1996; Keysar, Barr, & Horton, 1998), but there are at least two factors to consider in comparing the different findings. First, the contrasting studies are focused on different features of language. The phenomenon of interest in our experiment was the rapid deployment of grammar such as definiteness, deixis, modality, or voice during the course of interaction. The logic of these grammatical phenomena is not transparent to language users; it is only apparent after linguistic analysis. By contrast, people in Keysar's studies are dealing with lexical, not grammatical, items—comparing, for example, a "large" and "small" candle. Such items have meaning that is consciously accessible to users, and it is not surprising that their use is consciously evaluated. Second, participants in the experiments conducted by Keysar and colleagues interact in a situation that is not wholly shared. The settings for the experiments have features that are not accessible to both participants; speakers are aware that what they see is not the same as what their conversational partner sees. When people do not share context in a communicative situation, it can be expected that they will need to make a more conscious effort to make their perspective available to, or understand the perspective of, their conversational partner. This phenomenon of more effortful framing due to features of unshared context is one of the features characteristic of written as opposed to oral language register. Though the exchange is oral, Keysar and colleagues' experimental paradigm moves participants' linguistic production closer to the written end of the oral–written continuum.

Our experiment went a certain distance toward demonstrating that some features of the conversational interaction receive systematic, reliable, conscious episodic representation. But our methodology, dependent as it is on conscious verbal recall, is in principle unsuited for teasing out the speaker's implicit, subconscious mental models of the hearer's constantly shifting belief and intentional states. If grammar is a highly automated processing device, then the models of the interlocutor's mind associated with each grammatical device are constructed automatically and thus remain largely implicit. As philosophers, linguists, or discourse analysts, we may argue—given suggestive data that cannot be explained otherwise—that such mental models *must* exist. But the modal force of that "must" remains that of a hypothesis (Hanson, 1958).

Teasing out implicit mental models of whatever kind probably cannot be accomplished with conscious verbal recall. A less direct methodology, relying on more subtle retrieval clues—such as semantic priming—may be necessary. The phenomenon we seek is as elusive as it is pervasive. The mental representation of other minds in a social, communicating species is perhaps like the air we breathe—we hardly notice it, in spite of its extreme adaptive value, because it is so hopelessly ubiquitous.

REFERENCES

Anderson, A. S., Garrod, C., & Sanford, A. J. (1983). The accessibility of pronominal antecedents as a function of episodic shift in narrative text. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 35*(3-A), 427–440.

Blumstein, S. E., Milberg, W., & Shrier, R. (1982). Semantic processing in aphasia: Evidence from an auditory lexical decision task. *Brain and Language, 14*(2), 305–315.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Chafe, W. (1997). Polyphonic topic development. In T. Givón (Ed.), *Conversation: Cognitive, communicative, and social perspectives* (pp. 41–53). Amsterdam: Benjamins.

Coates, J. (1997). The construction of collaborative floor in women's friendly talk. In T. Givón (Ed.), *Conversation: Cognitive, communicative, and social perspectives* (pp. 55–89). Amsterdam: Benjamins.

DeSchepper, B., & Treisman, A. M. (1996). A visual memory for novel shapes: Implicit coding without attention. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22,* 27–47.

Dickinson, C., & Givón, T. (1997). Memory and conversation. In T. Givón (Ed.), *Conversation: Cognitive, communicative, and social perspectives* (pp. 91–132). Amsterdam: Benjamins.

DuBois, J. (1987). The discourse basis of ergativity. *Language, 63*(4), 805–855.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2).

Ervin-Tripp, S., & Kuntay, A. (1997). The occasioning and structure of conversational stories. In T. Givón (Ed.), *Conversation: Cognitive, communicative, and social perspectives* (pp. 133–166). Amsterdam: Benjamins.

Fernández-Duque, D. (1999). *Automatic processing object identity, location, and valence information.* Unpublished doctoral dissertation, University of Oregon, Eugene.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language.* Hillsdale, NJ: Erlbaum.

Gernsbacher, M. A. (1990). *Language comprehension as structure building.* Hillsdale, NJ: Erlbaum.

Givón, T. (1979). *On understanding grammar.* New York: Academic Press.

Givón, T. (1989). *Mind, code and context: Essays in pragmatics.* Hillsdale, NJ: Erlbaum.

Givón, T. (1991). Serial verbs and the mental reality of "event": Grammatical vs. cognitive packaging. In E. Traugott & B. Heine (Eds.), *Approaches to grammaticalization* (pp. 81–127). Amsterdam: Benjamins.

Givón, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics, 30, 5–55.*

Givón, T. (1994). Coherence in text, coherence in mind. *Pragmatics and Cognition, 1*(2), 171–227.

Givón, T. (2001a). *Syntax: An introduction*. Amsterdam: Benjamins.

Givón, T. (2001b). Toward a neuro-cognitive interpretation of "context." *Pragmatics and Cognition, 9,* 2.

Givón, T. (2002). *Bio-linguistics*. Amsterdam: Benjamins.

Givón, T. (2005). *Context as other minds*. Amsterdam: Benjamins.

Goodwin, C. (1988, November). *Embedded context*. Paper presented at the annual meeting of the American Anthropological Association, Phoenix, AZ.

Goodwin, C. (1995). The negotiation of coherence within conversation. In M. A. Gernsbacher & T. Givón (Eds.), *Coherence in spontaneous text* (Typological Studies in Language, Vol. 31, pp. 117–137). Amsterdam: Benjamins.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic Press.

Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge, UK: Cambridge University Press.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*(1), 91–117.

Kahneman, D., & Treisman, A. M. (1984). Changing views of attention and automaticity. In R. Parasuraman (Ed.), *Varieties of attention* (pp. 29–61). London: Academic Press.

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science, 7*(2), 46-50.

Kintsch, W. (1992). How readers construct situation models for stories: The role of syntactic cues and causal inference. In A. F. Healy, S. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes* (pp. 261–278). Hillsdale, NJ: Erlbaum.

Neville, H., Mills, D. L., & Lawson, D. S. (1992). Fractionating language: Different neural systems with different sensitive period. *Cerebral Cortex, 2*(3), 244–258.

Nissen, M. J., & Bulleme, P. (1986). *Attention requirements of learning: Evidence from performance measures*. Unpublished manuscript, University of Minnesota, Minneapolis.

Posner, M. I., & Peterson, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25–42.

Posner, M. I., & Snyder, C. R. R. (1974). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Sachs, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696–735.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search and attention. *Psychological Review, 84*(1), 1–66.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge, UK: Cambridge University Press.

Sieroff, E., & Posner, M. I. (1988). Cuing spatial attention during processing of words and letter strings of normals. *Cognitive Neuropsychology, 5,* 451–472.

Trabasso, T., Suh, S., & Payton, P. (1995). Explanatory coherence in understanding and talking about event. In M. A. Gernsbacher & T. Givón (Eds.), *Coherence in spontaneous text* (Typological Studies in Language, Vol. 31, pp. 189–214). Amsterdam: Benjamins.

Treisman, A. M., & DeSchepper, B. (1996). Object tokens, attention and visual memory. In T. Inui & J. McClelland (Eds.), *Attention and performance: Vol. 16. Information integration in perception and communication* (pp. 15–46). Cambridge, MA: MIT Press.

Wilkes-Gibbs, D., & Clark, H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language, 31*(2), 183–194.

15

Conceptual Alignment in Conversation

MICHAEL F. SCHOBER

Conversation is a primary site for inferring what is going on in other minds. The words and sentences people say, as well as how they say them, give strong evidence about their communicative intentions, any other intentions they may not intend to communicate, and about their mental states more generally. Just how reliable is this evidence? When people believe they have understood what is going on in each other's minds, how often are they right?

Views on these questions vary substantially, both on the street and among scholars, in part because they raise even more fundamental questions about the nature of language and about the nature of mental states. Does language simply encode mental states transparently, or are mental states not the sort of thing that language can embody directly? How much of thinking is encodable in the discrete and linear forms that language requires, and how much is ineffable and inarticulable? Do speakers of a particular language share identical context-free meanings of words, or do individuals differ in their semantic representations based on their personal experience of the world so that their personal meanings overlap only partially with other people's personal meanings?

Here I focus on one question among this larger set: To what extent do conversational partners mean exactly the same thing when they use the same words? Now, obviously conversational partners sometimes fail to understand each other's references. When Jennifer says to Don, "Look at that man!" Don may at first fail to recognize which man she is pointing out. To resolve this reference failure, Jennifer and Don can use any of the well-documented conversational techniques that speakers and addressees have for just this sort of problem (see Clark, 1996; Clark & Wilkes-Gibbs, 1986; Schegloff, 1988, among many others). Don can ask "Which man?" or Jennifer can clarify the reference without explicit prompting when Don doesn't respond appropriately ("You know, the one with the unusual hair"). They can take several conversational turns to agree that they have understood each other well enough for their current purposes—that they have *grounded* the reference, to use Clark's term.

The issue I raise here, although related, is a bit different. To what extent is Don's conceptualization of "man" (either in general or at the moment of Jennifer's reference) the same as Jennifer's? The fact that Jennifer and Don can agree that they are talking about the same man—that they have successfully used the word "man" for referring—does not guarantee that their mental representations for what counts as a man are identical. And this may be independent of how many turns it takes for them to establish the shared reference.

I propose that linguistic coordination is deceptive, and that it can mask undetected—even unsuspected—conceptual misalignment. Conceptual misalignments occur when interlocutors' mental representations have different content, for example when one person's category includes different exemplars or prototypes than the other's, or when one person's concept is elaborated in greater detail than the other's. A misalignment will be undetected if it doesn't lead to a reference failure, because the interlocutors will never think that they need to uncover it. So, for example, Jennifer and Don may happen to have different intuitions about membership in the category "man": different notions about when a teenager is old enough to be called a man, whether a presurgery female-to-male transgendered individual counts as a man, or whether inanimate statuary allows a human reference. But this won't come into play for many occasions of referring to "that man," because the misalignment isn't relevant for those circumstances. Only when circumstances are relevant to the misalignment, and when the conversationalists' desire for precision is high enough, might a misalignment be detected.

This proposal runs counter to the kind of argument advanced by Pickering and Garrod (2004) that linguistic coordination in dialogue automatically leads to conceptual alignment. Under that view, listeners' comprehension processes parallel speakers' production processes, and so when speakers use words, listeners' own (presumably identical) representations of the meanings of those words are automatically activated. Conceptual alignment is thus a natural feature of language use in dialogue.

I think the story is more complicated. I am not convinced that the evidence for automatic alignment is as compelling as Pickering and Garrod argue (see Schober, 2004, and the other replies for details). I am not convinced that people's conceptual misalignments are always detected, nor that people are particularly good at judging how aligned they are. And I suspect that undetected conceptual misalignments can have more serious consequences than one might at first think. Consider arguments—from spousal to international—about politically charged issues like abortion and euthanasia. If interlocutors don't know that they are considering quite different instances even though they are using the same term, the direction of argumentation and the resulting conflicts can escalate dangerously (see Schober, 1998b). Or consider whether job applicants or students taking standardized tests actually interpret the words in the questions and instructions in the same way as those administering job interviews and designing tests. If they do not, then school admissions and employment prospects could be judged on criteria other than those that test designers and interviewers intend, and in ways that exacerbate (or even create?) societal bias.

My main piece of evidence for this proposal comes from studies on how people comprehend ordinary terms in standardized survey interviews about facts and behaviors (e.g., Belson, 1981, 1986; Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad, & Fricker, 2004; Suessbrick, Schober, & Conrad, 2000). Such studies provide a useful entry into conceptual alignment in several ways.

STANDARDIZED SURVEYS

What makes surveys a good setting for studying conceptual alignment? First, unlike more artificial laboratory settings, a survey interview is a compelling real-world conversational arena in which conceptual misalignment can have serious consequences. Influential public policy and economic choices are made based on findings collected entirely from survey interactions in the U.S. Census (e.g., political redistricting and corporate moves based on changing demographics) and the Current Population Survey (e.g., federal and Wall Street decisions based on changes in the unemployment index). In this sort of situation, if survey respondents regularly conceptualize words in questions differently from the survey designers, the results could be disastrous.

Second, conceptual misalignment can be measured in surveys about facts and behaviors. Designers of major U.S. surveys have well-developed definitions for the terms used in those surveys, with extensive detail on what should be counted as (for example) a "job" or a "household member" for purposes of the survey. If we also have evidence about respondents' actual circumstances (what their job or living situation is like), we can then use respondents' answers to survey questions to assess whether their conceptualizations match the survey designers'. For example, if a respondent reports that her college-age child who lives in a dorm most of the year is a member of her household, we have evidence that her conceptualization of "household member" differs from that of survey sponsors who do not include children away at school.

Third, the predominant approach advocated for carrying out large-scale surveys is strictly standardized interviewing, in which the interpretation of terms in questions is left entirely up to respondents (e.g., Fowler & Mangione, 1990). If respondents request clarification of terms, interviewers are instructed to probe nondirectively: to simply repeat the question, repeat the response alternatives, or say something like "whatever it means to you." (The logic behind this is that if some interviewers defined terms for respondents and others didn't, the survey stimuli would no longer be uniform.) The fact that clarification in such surveys is forbidden means that we can assess respondents' untutored conceptualizations "in the wild."

Note that standardized interviewing techniques that leave the interpretation of words up to respondents embody precisely the same set of assumptions about the nature of linguistic meaning and about the problem of other minds that I am questioning. These assumptions include that wording and meaning are the same thing (see Schober, 1998a; Schober & Conrad, 2002; Suchman & Jordan, 1990), that conceptualizations within a community are functionally identical (see Pickering & Garrod, 2004), and that survey pretesting adequately uncovers most conceptual misalignments.

The studies described here extend findings from more controlled laboratory studies on how respondents answer questions about fictional scenarios (Schober & Conrad, 1997; Schober, Conrad, & Fricker, 2004). Here respondents answer questions about their actual life circumstances. As we did not have access to respondents' actual circumstances through official records or diaries (which can themselves be inaccurate), we measured conceptual alignment using two postsurvey measures.

In both studies, after answering the survey questions respondents answered more questions that allow us to infer their interpretations. One kind of question asked respondents to explain what they had included in their answers, either by listing examples or by answering multiple-choice questions about their interpretations. Another kind of question asked respondents to answer the same survey questions again, given standard definitions of the terms in the questions. If responses changed, then this suggests not only that respondents' initial conceptualizations were misaligned with the survey designers' but also that the misalignment was consequential enough that it had affected the survey data.

STUDIES

In one study (Conrad & Schober, 2000), 227 respondents from a nationally representative sample of residential households with telephones in the continental United States were interviewed by 10 professional telephone interviewers. Each respondent, after agreeing to participate in two different interviews 1 week apart, was asked 10 questions from ongoing U.S. government surveys. Five questions about housing, each requiring a numerical response, were taken from the Consumer Price Index Housing survey; respondents were asked how many bedrooms, full bathrooms, half-bathrooms, and other rooms other than bedrooms and bathrooms were in their home, as well as how many people lived in their home. Five questions about purchases, each requiring a yes/no response, were taken from the Current Point of Purchase Survey; respondents were asked if during a particular period (the past month, the past year) they had had any purchases or expenses for moving, telephones or telephone accessories, inside home maintenance or repair services, household furniture, and whiskey or other alcohol for home use. If respondents answered "yes" to any of these purchase questions, they were asked to list what those purchases had been. Interpretation of the terms in the questions was left entirely up to respondents; if respondents asked for clarification, interviewers politely repeated the question or otherwise refused to clarify.

One week later, the same respondents were interviewed again by 10 different interviewers. Exactly the same questions were asked the second time. This time, half the interviews were again carried out with no clarification of survey terms, following the norms of strict standardization. The other half were carried out by interviewers trained to make sure that respondents had interpreted the survey terms as the survey designers had intended. These interviewers were licensed, after initially reading the question exactly as worded, to present any or all of the complete official definitions for terms in the questions, in their own words or reading the scripted definition. For example, when respondents in this sort of interview were asked how many people lived in their home, interviewers could present any or all of this definition:

> A person is considered to be living in a housing unit even if the person is not present at the time of the survey. Live-in servants or other employees, lodgers, and members of the household temporarily away from the unit on business or vacation are included in the count. Do NOT count any people who would normally consider this their (legal) address but who are LIVING away on business, in the armed forces, or attending school (such as boarding school or college). Do NOT count overnight lodgers, guests and visitors. Do NOT count day employees who live elsewhere.

Interviewers could present clarification either when respondents explicitly requested it or also if they got the sense that presenting the definition or asking further questions would help a respondent interpret the question as intended.[1] Not surprisingly, these more collaborative interviews took longer to implement than the strictly standardized interviews.

Conceptual alignment between the respondents and survey designers was measured in two ways. First, the extent to which responses in the second interview differed from those in the first interview was assessed. If responses changed more when the second interviews included clarification, this would suggest that the initial interviews involved notable conceptual misalignment. Second, all the purchases that respondents had listed in both interviews were coded for whether they matched the requirements of the official definitions. This allowed assessment of whether changes in responses resulted from improved conceptual alignment.

The results showed that respondents had substantial undetected conceptual misalignment in the strictly standardized interviews. Twenty-two percent of responses changed when the second interview involved clarification, twice as many as the 11% response change when the second interview did not allow clarification.[2] Moreover, the response change in the interviews that allowed clarification did indeed reflect improved conceptual alignment. Among the purchases that respondents listed in the second (collaborative) interview, they correctly included items they had incorrectly excluded in the initial standardized interview: 90% of the purchases that were now listed in the interviews with clarification but that had not been included in the first interview matched the survey definitions. Similarly, respondents correctly excluded what they had erroneously included before: 89% of the purchases listed in the first interview that were now excluded with clarification had involved conceptual misalignments. (In contrast, for respondents who never received clarification, their listed purchases in the second interview were no more likely to reflect conceptual alignment than those in the first interview.)

Two more points about conceptual misalignment emerge from the data. First, the nature of the misalignments differed for different terms. For some of the question terms, the misalignments revolved around a few features of the official definitions. For example, for "moving" purchases, the vast majority of misalignments had to do with the fact that many respondents included personal expenses like van rental and hotel stays as moving expenses, while the definition excluded them. This would suggest not only that the misalignment is easily resolved (a simple rewording of the question should relieve it) but also that the population of respondents had substantially overlapping conceptions of what counts as moving. That is, despite the misalignment with the official survey definition, most respondents seemed to be conceptually aligned with one another. In contrast, for other terms (like "inside home maintenance or repair services") respondents were not at all aligned either with the survey designers or with one another; the substantial number of misalignments was distributed throughout all features of the official definitions.

Second, despite the substantial conceptual misalignments, respondents gave almost no evidence of being aware of them. Respondents in the interviews that allowed clarification were specifically instructed that this interview would allow clarification and that they should request it if they had even the slightest doubts about what any terms in the questions meant. Yet, in a subsample of 35 transcribed interviews, respondents explicitly requested clarification for only 6 of 165 questions—4% of the time.

A second study (Suessbrick, Schober, & Conrad, 2000) examined conceptual misalignment in a full-length survey that included both behavioral and opinion questions. The survey was the Tobacco Use Supplement to the Current Population Survey, which measures people's knowledge of and opinions toward smoking and tobacco use, and changes in their use over time. The survey is sponsored by the National Cancer Institute and administered by Census Bureau telephone interviewers; respondents answer from 12 to 36 questions, depending on their path through the survey.

The terms in the survey questions do not, on the surface, seem in any way ambiguous. Behavioral questions include "Have you smoked at least 100 cigarettes in your entire life?" and "Have you ever stopped smoking for one day or longer because you were trying to quit smoking?" Attitude questions include "In restaurants, do you think that smoking should be allowed in all areas, allowed in some areas, or not allowed at all?"

Official definitions existed for some of the questions, as in "*Past 12 months* means 12 months from today, NOT from the first of the month and not just the last calendar year." We supplemented them with additional definitions for undefined survey concepts, as in "By *smoked* we mean any puffs on any cigarettes, whether or not you inhaled AND whether or not you finished them."

Fifty-three respondents were interviewed by telephone in a laboratory by 10 experienced Census Bureau interviewers implementing strictly standardized procedures. No clarification of terms in the survey was ever given, even if respondents asked for it; respondents were required to interpret questions for themselves. After the telephone survey, respondents filled out two extended questionnaires. In the conceptualization questionnaire, they were asked multiple-choice questions probing how they had interpreted terms in each survey question they had answered in the telephone survey; for example, respondents were asked whether in answering the survey question "Have you smoked at least 100 cigarettes in your entire life?" they had interpreted "cigarettes" as including manufactured cigarettes, hand-rolled cigarettes, marijuana cigarettes, cigars, clove cigarettes, or something else (respondents were to pick all that applied). In the self-administered re-interview questionnaire, they answered the same survey questions as in the telephone interview; this time, half the respondents were to answer the question following a standard definition like the following (for "Have you smoked at least 100 cigarettes in your entire life?"):

> We want you to include any puffs on any cigarettes, whether or not you inhaled AND whether or not you finished them. We want you to include hand-rolled cigarettes as well as manufactured ones, and tobacco cigarettes with additives like cloves. We DON'T want you to include cigars or nontobacco cigarettes, like marijuana cigarettes.

The results suggest that respondents' conceptualizations were frequently misaligned with the official definitions and with the other respondents' interpretations. On the multiple-choice questionnaire, an average of fewer than 50% of responses perfectly matched the official definitions. Conceptual alignment was poor both for opinion questions (46% matched with official definitions) and for behavioral questions (33% matched). And this wasn't simply because the official definitions failed to reflect uniform interpretations held by the population of respondents; respondents' interpretations were not uniform but distributed among multiple interpretations. For the 37 concepts in questions answered by all respondents, on average only 51% of respondents endorsed the majority interpretation. For example, 46% of the respondents reported they had only considered "smoking" to include puffs that were inhaled; 54% reported that they included all puffs, whether or not they were inhaled. Similarly, 23% of respondents reported that they had interpreted "cigarettes" as including only cigarettes that they had finished smoking; 23% reported including both cigarettes they had finished or that they had only partly smoked; and 54% reported including any cigarettes they had taken even one puff of.

So, not only did respondents' self-reported interpretations match the survey definitions less than 50% of the time, they also did not match each other's. This conceptual variability is, of course, only interesting in practical terms if it affects the responses—and it did. Without a standard definition on the re-interview questionnaire, respondents changed fewer than 6% of their answers. With a standard definition, respondents changed 10% of their answers to the behavior questions and 16% of their answers to the opinion questions.

As in the previous study, respondents seemed quite unaware of their conceptual misalignments—or at least unwilling to do anything to resolve them during the initial telephone interviews. In two additional conditions of the experiment, 51 additional respondents were trained to ask for clarification during the telephone interview if they ever felt they needed it; in one condition interviewers were also trained to offer clarification if they believed the respondent needed it. Despite these instructions, only one respondent ever asked for clarification, and this only once. And despite extensive training on the potential for respondents' conceptual misalignments, interviewers almost never offered clarification. All in all, it didn't seem to occur to respondents or interviewers that respondents could conceive of terms in these questions differently than intended.

Conceptual misalignment can have consequences beyond undetected misunderstanding of a current reference; it can affect subsequent interaction as well. In this study, it led some respondents down the wrong line of questioning in the survey. The first question in the survey is "Have you smoked at least 100 cigarettes in your entire life?"; depending on whether respondents answer yes or no, they are led down the path of being a smoker or a nonsmoker. Among the 78 respondents who were given standard definitions in the re-interview questionnaire, 8 of them—10%—changed their responses to this question. This means that 10% of the respondents had been asked the wrong questions or hadn't been asked the right questions, and at a rate that could actually affect the substantive survey data.

QUESTIONS

These findings raise as many questions as they answer. Obviously they raise practical questions for survey researchers about how best to assure uniform interpretation of the survey questions they ask; if different respondents don't recognize that their conceptions may differ from survey designers', then standardized interviewing may unintentionally *increase* measurement error (Houtkoop-Steenstra, 2000; Schober & Conrad, 1997, 2002; Suchman & Jordan, 1990). These questions are serious not only for telephone interviews but also for face-to-face interviews (see Conrad, Schober, & Dijkstra, 2004), Web surveys (Schober, Conrad, & Bloom, 2000), and paper-and-pencil surveys. My colleagues and I (Schober, 1998a, 1999; Schober, Conrad, & Fricker, 2004) would argue that they raise the same questions for anyone who uses standardized wording in settings that don't allow clarification: designers of standardized educational tests, intelligence tests, tax forms, computer interfaces, experiment instructions—even newspaper reporters, novelists, and writers of scholarly research articles.

These issues are fundamental ones for investigating the problem of other minds. To what extent do other people's conceptualizations match our own? And under what circumstances do any conceptual misalignments matter? The data from these studies suggest at least one part of an answer: people's conceptualizations can match less often than common wisdom and prevailing theories suggest. And undetected conceptual misalignment is not benign precisely in those situations where there are consequences to categorizing events or objects differently than one's interlocutor, as in surveys where the accuracy of the data matters.

The data raise yet more fundamental questions. To what extent do people have stable and inflexible definitions of words? Just as people's judgments of category membership seem more flexible than classical theories suggest (see reviews in Margolis & Laurence, 1999), perhaps people's semantic systems are less stable than has been assumed. Our survey respondents seemed perfectly willing and able to conceive of "smoking" through an alternate lens when it was required of them. And certainly the evidence from other contexts of language use is that people can use the same words quite variably to refer to different entities. Is there such a thing as a context-free core for each concept, rigid across circumstances? Do different people have different cores, or can even the same people shift meanings as circumstances require?

The data also raise the question of whether and how accurately people build models of their conversational partners, and the role such models play in online processing (see Schober & Brennan, 2003). Our survey respondents did not seem particularly aware of the potential for conceptual misalignment. Why not? Although we can't rule out the possibility that some of our survey respondents simply didn't care enough to try hard to resolve misalignment, this can't be all there is to it; when given standard definitions, respondents willingly changed their answers to fit those requirements. I believe the issue reflects what Herb Clark and I have called the *presumption of interpretability* (Clark & Schober, 1991). At some level, respondents presumed that their concepts (which were, after all, quite ordinary—bedrooms, smoking, living in a home) matched survey designers'. Just as in spontaneous conversation, survey respondents presume that their partners have designed utterances with them in mind, and interpret accordingly. The addressee presumes that if a speaker intends something by an utterance other than what the addressee initially takes it to mean, then it is the responsibility of the speaker to clarify it; otherwise, the addressee is licensed to go with his or her initial interpretation.

Another question: Even if people don't consciously recognize conceptual misalignments, do they provide any signs of processing difficulty that would show implicit knowledge? The evidence from surveys thus far is that even when respondents don't think the misalignments are problematic enough to ask for clarification, they can still "leak" clues of it in the way they talk. Respondents are more likely to use "um" and "uh," restart their utterances, or describe their circumstances rather than answering the survey question in situations of potential conceptual misalignment (Schober & Bloom, 2004). Various paralinguistic and facial cues seem to indicate people's processing difficulty, whether or not the cues are intended as communicative signals (Clark, 1994, 1996), and listeners seem to make use of speakers' clues in their moment-by-moment comprehension processes (e.g., Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Brennan & Schober, 2001; Fox Tree, 1995, 2001). On the other hand, some percentage of conceptual misalignments seem entirely undetected: In our survey study (Schober & Bloom, 2004) sometimes survey respondents provided no clues of trouble at all, answering quickly and confidently even when conceptually misaligned.

Despite all the questions, what is clear is that linguistic alignment does not guarantee conceptual alignment and that undetected misalignments can have serious consequences. This is consistent with the long skeptical tradition arguing that we cannot be certain that the contents of other people's minds are identical to our own. The trouble—for researchers and for interlocutors—is that it remains unclear exactly when undetected misalignments are having consequences worth worrying about. For Jennifer and Don referring to "that man," the consequences are probably negligible; for high-stakes surveys the consequences can be serious. It is unknown how well conceptualizations match, and how often consequential misalignments are undetected, in other discourse settings. The implication is that when we are in high-stakes situations, like international negotiations and marital disputes, we might be wise to check the extent to which we understand other people's words in the same way.

ACKNOWLEDGMENTS

This work was funded in part by National Science Foundation Grant Nos. SBR-9730140 and ISS-0081550 and by the U.S. Bureau of Labor Statistics.

NOTES

1. Interviewers varied substantially in the strategies they took, as one might expect if they were adapting to the different conversational needs of different respondents. As more controlled laboratory studies (e.g., Lind, Schober, & Conrad, 2000; Schober, Conrad, & Fricker, 2004) have confirmed, what is crucial is whether respondents get the relevant piece of a definition, not whether they receive it at their request.
2. An 11% rate is not unusual for re-interviews in large-scale surveys, and no doubt reflects multiple sources: memory errors at one time or another, actual changes in life circumstances—and most interestingly for our purposes, possible instability of conceptualizations within individuals.

REFERENCES

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science, 15*, 578–582.

Belson, W. A. (1981). *The design and understanding of survey questions*. Aldershot, UK: Gower.

Belson, W. A. (1986). *Validity in survey research*. Aldershot, UK: Gower.

Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language, 44*, 274–296.

Clark, H. H. (1994). Managing problems in speaking. *Speech Communication, 15*, 243–250.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Clark, H. H., & Schober, M. F. (1991). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 15–48). New York: Russell Sage Foundation.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1–39.

Conrad, F. G., & Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64*, 1–28.

Conrad, F. G., Schober, M. F., & Dijkstra, W. (2004). Nonverbal cues of respondents' need for clarification in survey interviews. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error.* Newbury Park, CA: Sage.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language, 34*, 709–738.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory and Cognition, 29,* 320–326.

Houtkoop-Steenstra, H. (2000). *Interaction and the standardized survey interview: The living questionnaire*. Cambridge, UK: Cambridge University Press.

Lind, L. H., Schober, M. F., & Conrad, F. G. (2001). Clarifying question meaning in a web-based survey. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Margolis, E., & Laurence, S. (Eds.). (1999). *Concepts: Core readings*. Cambridge, MA: MIT Press.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190.

Schegloff, E. A. (1988). Discourse as an interactional achievement: II. An exercise in conversation analysis. In D. Tannen (Ed.), *Linguistics in context: Connecting observation and understanding* (pp. 135–158). Norwood, NJ: Ablex.

Schober, M. F. (1998a). Conversational evidence for rethinking meaning. *Social Research*, *65*(3), 511–534.

Schober, M. F. (1998b). Different kinds of conversational perspective-taking. In S. R. Fussell & R. J. Kreuz (Eds.), *Social and cognitive psychological approaches to interpersonal communication* (pp. 145–174). Mahwah, NJ: Erlbaum.

Schober, M. F. (1999). Making sense of questions: An interactional approach. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research* (pp. 77–93). New York: Wiley.

Schober, M. F. (2004). Just how aligned are interlocutors' representations? Commentary on Pickering and Garrod. *Behavioral and Brain Sciences*, *27*, 209–210.

Schober, M. F., & Bloom, J. E. (2004). Discourse cues that respondents have misunderstood survey questions. *Discourse Processes, 38,* 287–308.

Schober, M. F., & Brennan, S. E. (2003). Processes of interactive spoken discourse: The role of the partner. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 123–164). Mahwah, NJ: Erlbaum.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, *60,* 576–602.

Schober, M. F., & Conrad, F. G. (2002). A collaborative view of standardized survey interviews. In D. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 67–94). New York: Wiley.

Schober, M. F., Conrad, F. G., & Bloom, J. E. (2000). Clarifying word meanings in computer-administered survey interviews. In L. R. Gleitman & A. K. Joshi

(Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 447–452). Mahwah, NJ: Erlbaum.

Schober, M. F., Conrad, F. G., & Fricker, S. S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology, 18*, 169–188.

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association, 85*, 232–253.

Suessbrick, A. L., Schober, M. F., & Conrad, F. G. (2000). Different respondents interpret ordinary questions quite differently. In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 907–912). Alexandria, VA: American Statistical Association.

16

On the Inherent Ambiguity of Traits and Other Mental Concepts

JAMES S. ULEMAN

The attribution of personality traits to other people is ubiquitous. What traits refer to and the processes by which they are inferred have been researched by personality and social psychologists for well over half a century (e.g., Asch, 1946; Heider, 1944). Yet, we have only the most rudimentary ideas about what trait terms refer to, how they are inferred, and what affects these meanings and inferences. This chapter explores two ideas that have been relatively neglected but are central to understanding the meaning of trait terms: (1) that most traits refer to others' minds—their goals, beliefs, desires, intentions, fears, aspirations, etc.—and (2) that trait and other mental terms are inherently ambiguous.

Classic attribution theories of trait inference address the kinds of information that promote inferring causes (traits or dispositions) about actors rather than situations. Jones and Davis (1965) focused on intentional actions, noting the importance of their social desirability and unique effects for inferring something about the actor. But they told us little about how intentional actions differ from mere (accidental) behaviors (see Malle, 2005). Kelley (1967) pointed to the covariation (over multiple observations) of behaviors' effects with the presence of the actor, rather than the behaviors' object or the situation, as the basis for inferring that something about the actor caused the effect. But he offered few suggestions on how we identify precisely what it is about the actor that caused the effect. Thus, we infer "it's something about Mary," but do not know what that "something" is. Reeder and Brewer (1979) suggested several schemata that govern inferences of traits in the ability and morality domains. But none of these theories or related research offers fine-grained characterizations of the kind of information used to produce *particular* trait inferences.

Classic personality theories and research, which treat traits as invariant properties of people that do not depend on their situations or circumstances, do not advance this problem either, because their basic data are the very inferences, observations, and self-reports we are trying to explain. In addition, Wright and Mischel (1987) provide strong evidence that traits are essentially conditional and depend for their meaning on particular contexts, either explicitly described or implicitly understood. That is, traits do exist, but it makes no sense to treat trait terms as if they have meaning in isolation, out of context.

To complicate matters further, research over the past 20 years has shown that unconscious processes, including those that produce trait inferences, are at least as important as the conscious processes of person perception to which classic theories typically appeal (e.g., Hassin, Uleman, & Bargh, 2005; Nisbett & Ross, 1980; Uleman & Bargh, 1989). Furthermore, a small but growing body of research shows directly that conscious and unconscious trait inferences differ systematically from each other (Ham, 2004; Todorov, Gonzalez, Uleman, & Thaden, 2004; Uleman, 1999; Zárate, Uleman, & Voils, 2001).

So, it may be timely to speculate about what trait terms refer to and how their meaning relates to the psychological processes that produce trait inferences. I focus on the idea that trait terms are ambiguous (as are other mental concepts). I review some evidence for this claim and note a variety of ambiguities in traits' meanings. Then I mention two classes of knowledge structures in which trait concepts are important participants, and urge future research on them. These are (1) theories of mind and (2) abstract mental concepts based on metaphors from our bodily experience, our embodied cognitions.

TWO VIEWS OF PERSONALITY TRAITS

When you describe someone as "hard" or "warm" or "blue," what might you mean? What is the nature of the concept you are using? More specifically, do these terms refer to objective properties of the person that we could measure if we had the proper instruments, in the same way you can measure a piece of glass's hardness (on the Mohs scale), or its temperature (in degrees Celsius), or its color (in angstroms)? Or do these terms' meanings depend on the context in which they are used, with virtually every context imparting a different meaning? How does a hard (energetic) worker resemble a hard (stingy) paymaster, or a judge who is hard on (punitive toward) criminals, or a hard (obdurate, stubborn) opponent? The "warmth" of a lover and the "warmth" of a parent should be quite different in several ways. One's response to friends who are "a little blue" depends on whether you think that they feel sad or ribald. Descriptions of people may be inherently more ambiguous[1] than descriptions of objects. In what ways might this be true?

As a first approximation, there are two views of the meaning of personality traits. The first assumes that these meanings are clear and relatively invariant; the second does not. In one variant of this first view, traits describe distinct properties of people that can be inferred from their behavior, their talk, and their reactions to particular situations. They constitute the basic level of the fuzzy categories of personality (John, Hampson, & Goldberg, 1991). In another variant, trait terms function as behavior categories with a graded structure in which some behaviors are more prototypical of a trait than others (e.g., Buss & Craik, 1981). This view that trait terms (in isolation) have clear meanings allows them to be related to one another in relatively invariant ways, which may explain why factor analyses of trait ratings consistently yield the "Big Five" global factors—extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience—each of which subsumes other traits in a hierarchical structure. Some (e.g., John, 1990) think this reflects semantic structure, whereas others (McCrae & Costa, 1999) give these factors causal status in their "five-factor model of personality." In either case, the meanings of traits are clear and unique, and defined by their relationships to other traits and behaviors.

This view is probably also dominant in social psychology. Research on person perception, beginning with Asch's (1946) classic studies, has emphasized how people use behavioral, situational, and trait information to form trait impressions. Associative theories based on Anderson and Bower's (1973) human associative memory (HAM) model remain popular, with connected nodes usually representing persons, traits, and behaviors. More generally, trait concepts are taken as invariant and unanalyzed primitives in theories of impression formation (e.g., Wyer & Lambert, 1994). When "a trait concept" is "activated," the implication is that there is an "it" waiting in memory to be activated and applied. Trait meanings are not selected from multiple possibilities or computed anew online. Even though Asch posited the context-dependent nature of trait meanings (i.e., senses), and this was clearly demonstrated almost 30 year later (Hamilton & Zanna, 1974), most research on person perception (see Gilbert, 1998) has largely ignored this context dependence.

The second view of personality trait terms is that they have multiple rather than single senses, and that context selects relevant senses, often without users' conscious awareness. There are two lines of research on personality that support this view. First, exploratory factor analyses of trait ratings of individual targets rated repeatedly over many days (rather than ratings of many targets rated once) do not yield the familiar "Big Five" factor structure for most targets. Nesselroade and Molenaar (1999) found that less than a third of their targets showed the familiar five-factor pattern of trait covariation over time, and Borkenau and Ostendorf (1998) put that figure at 10%. In arguing that traits' senses depend on context, Caprara and Cervone (2000, p. 79) put it this way: "The lexical items [traits] people use to describe personality can not be treated as if they are fixed elements of a periodic table."

Second, Wright and Mischel (1987) looked at the empirical determinants of camp counselors' judgments of how aggressive and withdrawn summer camp children were. They showed that these judgments reflected the children's behavior in highly circumscribed situations. For example, "aggressiveness" referred to behavior in specific situations that taxed the child's competencies to handle those situations (such as playing cooperatively with peers, or complying with adults' requests), not their behavior in all or even most situations, and not behavior in the same kinds of situation for every child. Wright and Mischel (1988) showed further that counselors understood the conditional nature of their trait judgments. Open-ended descriptions of children (rather than ratings on context-free traits) "systematically linked specific categories of conditions (e.g., aversive interpersonal events) to specific categories of social behavior (e.g., aggressive acts)" (p. 454). That is, when they labeled a child as "aggressive," they usually had contextual qualifiers in mind, such as "aggressive in that the child acts in X way when in Y contexts."

While personality researchers have been documenting the polysemy (see note 1) of trait terms in descriptions of people, social psychologists have been exploring the ambiguity of behavior descriptions. Trope (1986; Trope & Gaunt, 1999) provided an elegant model and experimental data to show that situational contexts affect the trait-relevant meanings of behaviors. Crying at a funeral is understood differently from crying at a wedding (and each has distinct implications for the person's personality). In addition, decades of research on priming and category accessibility in social cognition show that trait interpretations of behaviors can depend on contexts, including contexts that are logically irrelevant and even outside of awareness (see Higgins, 1996, for a review). Thus, the fictional Donald's mountain climbing and kayaking the rapids can be interpreted as adventurous or reckless. How you interpret it depends on which trait construct is most accessible when you hear about Donald (Higgins, Rholes, & Jones, 1977).

Social psychologists have not neglected the polysemy of trait terms, either. Considerable research has shown how people unknowingly exploit polysemy to maintain self-serving positive self-concepts. Most people believe they are better than average on a wide range of attributes. Dunning, Meyerowitz, and Holzberg (1989) showed that this "better-than-average" effect holds only for polysemous traits, that is, those with many behavioral indicators. Thus, most people report they are above-average on creativity, but they define it in different and self-serving ways (e.g., in the sense of musical or narrative or scientific creativity, depending on which talents they have). The "better-than-average" effect does not occur for unambiguous traits such as "punctual." Dunning and Cohen (1992) showed that on continuous traits and abilities high scorers set higher criteria for judging another's performance as "good" than low scorers. And Dunning and McElwee (1995) showed that people who rate themselves high on a trait (e.g., dominant) define it more in terms of positive behaviors than those who score low. Accordingly, experimentally varying the salience of positive versus negative behavioral senses of a trait affects people's self-ratings on that trait.

Thus, traits (and behaviors) have many senses. These multiple senses are not usually problematic because contexts disambiguate them. But are the cases cited above exceptional? Is there any evidence suggesting that traits (and behaviors) are more ambiguous than other concepts? Interestingly, there is. And this evidence points to a variety of types of ambiguities.

ARE TRAITS, MENTAL EVENTS, AND BEHAVIORS PARTICULARLY AMBIGUOUS CONCEPTS?

In reviewing evidence on this question, several kinds of ambiguity emerge. What I will call *boundary ambiguity* concerns how clearly entities are distinguished from their parts and their superordinate categories. Physical objects (such as apples or cars) are more distinct or bounded than activities (such as going to a movie or to a restaurant) or mental events (such as thinking or dreaming). Rips and Estin (1998) compared objects, scripted activities, and mental events in several ways. All of the comparisons suggested that objects are the most distinctive or bounded.

In their first study, participants generated parts of and superordinate categories for entities (specifically objects, activities, and mental events). Thus, for objects, an apple core is a *part of* an apple, and an apple is a *kind of* fruit. Similarly, for mental events, using logic is *part of* reasoning, and reasoning is a *kind of* thinking. For each entity (e.g., apple or reasoning), participants then judged whether the part (e.g., apple core or using logic, respectively) "is a kind of" the superordinate (e.g., fruit or thinking, respectively), and whether the entity (apple, reasoning) "is a part of" the superordinate (fruit, thinking). Although parts of objects were rarely seen as parts of their superordinate categories, parts of scripts and mental events often were. An apple core is not a kind of fruit, but using logic is a kind of thinking. That is, there was more homogeneity among the parts and wholes of scripts and mental events than objects. Furthermore, participants could name more distinctive properties for their objects than their scripts, and more distinctive properties for their scripts than their mental events. In addition, ratings of "boundedness" were highest for objects and lowest for mental events, and this boundedness predicted the other results better than other ratings. Thus metonymy (i.e., naming parts and wholes with the same term) is more common for scripted activities and mental events than for objects. Such terms are polysemous in terms of their referents' boundaries. Hampson, John, and Goldberg (1986) made a similar point about the way trait terms are structured, without providing explicit comparisons with other kinds of concepts.

Another way to demonstrate words' ambiguity is to compare their meanings in contexts with their meanings in isolation. The meanings of adjectives (such as traits) usually change when they are combined with nouns. Murphy and Andrew (1993) asked participants to list the opposites of 14 isolated adjectives and then list the opposites of 63 adjectives as used in adjective–noun phrases. Some phrases (e.g., cold water, bright light) were designed to elicit the same opposites as their adjectives alone, whereas others were not (e.g., cold facts, bright child). Indeed, participants' opposites of the former matched their opposites of the corresponding adjectives alone 66% of the time, whereas opposites of the latter matched them only 25% of the time. Some combinations appeared likely to elicit distinctly different meanings of the same adjectives (i.e., to represent homonyms; see note 1), whereas others (e.g., healthy appetite, healthy grip; fresh shirt, fresh water) appeared to be more polysemous. But the analyses do not distinguish between these two. In a second study, where adjective–noun phrases were selected at random from a large corpus of text, the mean match rate was only 51%. Two other studies asked for synonyms rather than antonyms, with similar results, including a mean match rate of only 44% for random adjective–noun phrases. Thus, adjectives' meanings show a high degree of what I will call *combinatorial ambiguity.*

Murphy and Andrew (1993) make a strong case that the meanings of many or most adjective–noun phrases are computed online rather than *prestored*. Traits are adjectives, and all 14 of the adjectives used in two of their studies could be used as traits. Indeed, illustrating the combinatorial ambiguity of trait adjectives, Kunda, Sinclair, and Griffin (1997) found that occupational stereotypes of actors changed the behavioral meanings of trait terms. For example, the combination of "aggressive" and "construction worker" implied physical aggression to their participants, whereas "aggressive lawyer" implied verbal aggression.

Many traits are also ambiguous in the sense that they can refer to, and be implied by, a wide range of very different behaviors. In this sense, creative is a relatively ambiguous trait and punctual is a relatively unambiguous trait. Such *referential ambiguity* is clearly documented, and its implications are developed in the research by Dunning and his colleagues described above.

Finally (but not exhaustively), behaviors show *valence ambiguity* in that they often imply traits of opposite valence. "Donald spent a great amount of his time in search of what he liked to call excitement. He had already climbed Mt. McKinley, shot the Colorado rapids in a kayak, driven in a demolition derby, and piloted a jet-powered boat—without knowing very much about boats. He had risked injury, and even death, a number of times" can imply that Donald is either adventurous (positive) or reckless (negative) (Higgins et al., 1977). This same classic study and many subsequent ones have described Donald's other behavior in ways that can be interpreted as independent or aloof, and as persistent or stubborn. Trait inferences from behavior may even require resolution of behaviors' valence ambiguity more often than not, because Cruse (1965) found that evaluations of English trait terms have a bimodal distribution, making neutral traits rare.

In short, behavior descriptions and mental action terms are more ambiguous than terms for objects in at least one way, and trait and behavior descriptions are ambiguous in other ways too. Yet, the meanings of all of these words are readily and adequately disambiguated in most daily discourse. There appears to be no "best" or "true" interpretation of trait terms independent of the (syntactic, semantic, and pragmatic) context in which they are used.

Trait terms can be used to describe act frequencies for individuals (Buss & Craik, 1981) or the goals of those acts (Read, Jones, & Miller, 1990). Borkenau (1990) provided evidence that traits represent ideals rather than central tendencies, "conveying information on the aptitude of persons for [attaining] several [of their own] goals" (p. 394) rather than their typical behaviors. Thus, people may be "helpful" in the sense that they just gave assistance, or frequently give assistance, or chronically want to give assistance, or are capable of giving assistance. Of course, particular forms of assistance depend on the circumstances and the person's other characteristics, making the multiple meanings of *helpful* innumerable if not infinite.

People prefer trait descriptions over goal descriptions when they are forming impressions or predicting future behavior, but goals are preferred when memorizing someone's acts or empathizing with them (Hoffman, Mischel, & Baer, 1984; Hoffman, Mischel, & Mazze, 1981). Traits can describe people or merely their behaviors (Todorov & Uleman, 2002), or even become incidentally associated with other people and objects (Brown & Bassili, 2001; Skowonski, Carlston, Mae, & Crawford, 1998), depending on the conditions under which traits are initially inferred from behaviors. Traits can provide interpretations of ambiguous acts, so that traits and act descriptions are assimilated, as when Donald's acts are interpreted as reckless after the concept of reckless has been activated in an apparently unrelated context (Higgins et al., 1977). Or traits can provide standards or ideals against which ambiguous acts are judged, so that traits and acts are contrasted, as when first explicitly describing reckless Erik makes Donald seem not so reckless but adventurous instead (Stapel & Koomen, 2001), again depending on the other conditions under which Donald's traits are initially inferred. Traits can describe temporary states or enduring characteristics, depending on one's beliefs about traits' malleability (Dweck, 1999).

Of course, this does not mean that trait terms can mean anything. But it does mean that traits are polysemous and ambiguous in other multiple ways. As Murphy and Andrew (1993) suggest, most of these meanings are probably computed online. So, rather than trying to discover traits' "true meaning," it seems more productive to ask what classes of knowledge give traits their meanings in particular situations.

TWO INTERESTING CLASSES OF KNOWLEDGE THAT DISAMBIGUATE TRAIT TERMS

Virtually any knowledge can disambiguate trait terms in the right circumstance, but two classes of knowledge seem particularly interesting and relevant. One is knowledge of people's mental states, based on theories of mind. At least one anthropologist (D'Andrade, 1987) and many developmental psychologists (e.g., Wellman, 1990, p. 115) include traits as higher-order constructs in their treatments of theory of mind. Idson and Mischel (2001) showed that, even though simple trait terms are most common in undergraduates' descriptions of most people most of the time, when the targets are better known and important to the judges, they are less likely to use traits. Instead, they use more "cognitive-affective mediating units" (CAUs). These include affects (moods and feelings), beliefs, competencies, encodings (interpretations and categorizations), expectancies, goals, needs, and values—that is, components of theory of mind. This suggests that when you know someone well you reach behind the inherent polysemy of traits to the (often situated) mental states that underlie them. It also suggests that traits are theory-based concepts (Murphy, 2002) in which theory of mind plays a large part.

Several social psychologists have elaborated on the ways that theory of mind is central to how we perceive and understand others in trait terms. As Reeder (Reeder & Trafimow, Chapter 7, this volume; Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002), Read (1987; Read & Miller, Chapter 8, this volume), and Ames (Chapter 10, this volume; 2004) attest, the idea that theory of mind underlies most trait inferences is not novel. What is novel is my suggestion that the complexity and context sensitivity of any such theory must result in giving the apparently simple inferences from it (e.g., traits) many different meanings.

I use "theory" loosely, as many psychologists do. But if one takes "theory" seriously, one might ask what kind of rigorously predictive theory can accommodate the polysemy, ambiguity, and context sensitivity described above. Kunda and Thagard (1996; see also Thagard & Kunda, 1998) describe a parallel constraint satisfaction model that is computationally tractable and can simulate many well-known phenomena in person perception. These include shifts in the meanings of behaviors and traits as a function of contexts, judgments of explanations' coherence, and reasoning about one person by forming an analogy with someone else.

A second class of knowledge is described by Lakoff and Johnson (1999). They suggest that abstract concepts (such as time, causality, the mind, the self, and morality) are largely metaphorical and based on concrete experience. For example, time is like a material resource such as money. You can spend time, save time, owe time, waste time, invest time, etc. Spatial experience provides other metaphors for time, with the future "in front of you" and the past "behind you." We move through time, and time passes us by. These last two metaphors are especially interesting because they can produce conflicting interpretations. If someone wants to move next week's meeting scheduled for Wednesday "ahead two days," and you are "moving through time," you're likely to think they want to move the meeting to Friday. But if you think of time "moving past you," you're likely to think they want to move the meeting to Monday. Boroditsky and Ramscar (2002) showed that thinking about spatial experience one way or the other can prime which metaphor is used to interpret such temporally ambiguous messages and, importantly, that the process is asymmetric. Thinking about time in one way or the other does not prime interpretations of spatially ambiguous messages.

Lakoff and Johnson (1999) are not the only ones to suggest that abstract concepts, including mental ones such as *mind* and *self*, derive from experience-based metaphors (e.g., Abelson, 1986). But their elaboration of this idea is especially provocative because it shows that (1) several metaphors may be used to understand the same abstract concept, and (2) these may be inconsistent with one another and even contradictory, as in the example above. Traits are certainly abstract mental concepts, often understood metaphorically. An angry person may be like a bomb (about to blow up, with a short fuse) or a tea kettle (simmering, boiling over, letting off steam). A smart person's mind may be like a knife (sharp, able to cut to the core of an issue) or an encyclopedia ("containing" a lot of authoritative knowledge, serious, "heavy") or a trap (quick to grasp knowledge, slow to relinquish it).

This emphasis on thinking of metaphors as based in physical experience, rather than merely in symbolic (verbal) knowledge, is very compatible with recent work on "embodied cognition." This work suggests that, rather than being represented exclusively in an amodal symbol system, knowledge is also represented "as partial simulations of sensory, motor, and introspective states" (Barsalou, Niedenthal, Barbey, & Ruppert, 2003, p. 44). Evidence for embodied cognition comes from research in cognitive, social, and developmental psychology, and Barsalou's (1999) theory of perceptual symbol systems provides a theoretical framework for integrating it. Although there is no research yet on the role of embodied cognition in trait inference and trait interpretation, it is clear that embodied cognition plays a role in comprehending social acts, even when they are only presented verbally (Richardson, Spivey, Barsalou, & McRae, 2003). In two studies, Richardson and colleagues (2003) found that, as predicted, comprehension of short auditory sentences containing verbs associated with horizontal or vertical image schemata (e.g., *push* or *respect*, respectively) differentially affected the subsequent processing of visual stimuli along these two spatial axes. That is, the comprehension of many action verbs (e.g., *pull, hunt, show, float, walk*) entails activating spatial schemata.

In short, our theories and future research on person knowledge and how it is processed can only be enriched by taking more explicit account of the ambiguity of trait terms and other mental concepts, and the kinds of knowledge structures that help disambiguate them.

ACKNOWLEDGMENTS

Work on this chapter was supported by Grant No. MH-069842 from the National Institutes of Health to James S. Uleman. Portions of this chapter were presented at the 2004 Werkmeister Conference: Folk Concepts: Philosophical and Psychological Perspectives, Tallahassee, Florida, January 18, 2004. I wish to thank participants in the Other Minds Conference in Eugene, Oregon, and participants in the Werkmeister Conference in Tallahassee for their comments and suggestions, especially philosophers Jonathan Adler (CUNY Graduate Center), Lynne Rudder Baker (University of Massachusetts, Amherst), John Doris (University of California, Santa Cruz), and Kirk Ludwig (University of Florida, Gainesville).

NOTE

1. Words can be ambiguous in many ways, one of which is polysemy. Polysemy refers to the multiple senses that one word (lexeme) or lexical unit can take, all of which share some core meaning. Polysemy differs from homonymy, in which the same lexical form refers to different families of meanings (e.g., "bank" as a financial institution or the boundary of a river), which are often listed separately in a dictionary. Thus the "hard" trait examples seem to share an underlying meaning of imperviousness, whereas "blue" has two different meanings. See Cruse (1986) for a much more nuanced and interesting discussion.

REFERENCES

Abelson, R. P. (1986). Beliefs are like possessions. *Journal for the Theory of Social Behavior, 16*, 223–250.

Ames, D. R. (2004). Inside the mind-reader's toolkit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology, 87*, 340–353.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 1230–1240.

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory.* Washington, DC: Winston.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., & Ruppert, J. A. (2003). Social embodiment. *Psychology of Learning and Motivation*, *43*, 43–92.

Borkenau, P. (1990). Traits as ideal-based and goal-derived social categories. *Journal of Personality and Social Psychology*, *58*, 381–396.

Borkenau, P., & Ostendorf, F. (1998). The big five as states: How useful is the five-factor model to describe intraindividual variations over time? *Journal of Research in Personality*, *32*, 202–221.

Boroditsky, L., & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science*, *13*, 185–189.

Brown, R. D., & Bassili, J. N. (2001). Spontaneous trait associations and the case of the superstitious banana. *Journal of Experimental Social Psychology, 38,* 87–92.

Buss, D. M., & Craik, K. H. (1981). The act frequency analysis of interpersonal dispositions: Aloofness, gregariousness, dominance and submissiveness. *Journal of Personality*, *49*, 175–192.

Caprara, G. V., & Cervone, D. (2000). *Personality: Determinants, dynamics, and potentials*. Cambridge, UK: Cambridge University Press.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge, UK: Cambridge University Press.

Cruse, D. B. (1965). Social desirability scale values of personal concepts. *Journal of Applied Psychology*, *49*, 342–344.

D'Andrade, R. (1987). A folk model of the mind. In D. Holland & N. Quinn (Eds.), *Cultural models in language and thought* (pp. 112–148). Cambridge, UK: Cambridge University Press.

Dunning, D., & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology*, *63*, 341–355.

Dunning, D., & McElwee, R. O. (1995). Idiosyncratic trait definitions: Implications for self-description and social judgment. *Journal of Personality and Social Psychology*, *68*, 936–946.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions and self-serving assessments of ability. *Journal of Personality and Social Psychology*, *57*, 1082–1090.

Dweck, C. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press/Taylor & Francis.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 89–150). New York: McGraw-Hill.

Ham, J. (2004). *Bridging attribution and spontaneous inferences: Spontaneous and intentional components of dispositional and situational inferences*. Doctoral dissertation, Katholieke Universiteit Nijmegen,The Netherlands.

Hamilton, D. L., & Zanna, M. P. (1974). Context effects in impression formation: Changes in connotative meaning. *Journal of Personality and Social Psychology*, *29*, 649–654.

Hampson, S. E., John, O. P., & Goldberg, L. R. (1986). Category breadth and hierarchical structure in personality: Studies of asymmetries in judgments of trait implication. *Journal of Personality and Social Psychology*, *51*, 37–54.

Hassin, R. R., Uleman, J. S., & Bargh, J. A. (Eds.). (2005). *The new unconscious*. New York: Oxford University Press.

Heider, F. (1944). Social perception and phenomenal causality. *Psychological Review*, *51*, 358–374.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York: Guilford Press.

Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141–154.

Hoffman, C., Mischel, W., & Baer, J. S. (1984). Language and person cognition: Effects of communicative set on trait attribution. *Journal of Personality and Social Psychology*, *46*, 1029–1043.

Hoffman, C., Mischel, W., & Mazze, K. (1981). The role of purpose in the organization of information about behavior: Trait-based versus goal-based categories in person cognition. *Journal of Personality and Social Psychology*, *40*, 211–215.

Idson, L. C., & Mischel, W. (2001). The personality of familiar and significant people: The lay perceiver as a social-cognitive theorist. *Journal of Personality and Social Psychology*, *80*, 585–596.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). New York: Guilford Press.

John, O. P., Hampson, S. E., & Goldberg, L. R. (1991). The basic level in personality-trait hierarchies: Studies of trait use and accessibility in different contexts. *Journal of Personality and Social Psychology*, *60*, 348–361.

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. 220–266). New York: Academic Press.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192–238). Lincoln: University of Nebraska Press.

Kunda, Z., Sinclair, L., & Griffin, D. (1997). Equal ratings but separate meanings: Stereotypes and the construal of traits. *Journal of Personality and Social Psychology*, *72*, 720–734.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors. A parallel constraint satisfaction theory. *Psychological Review*, *103*, 284–308.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.

Malle, B. F. (2005). Folk theory of mind: Conceptual foundations of human social cognition. In R. R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 225–255). New York: Oxford University Press.

McCrae, R. R., & Costa, P. T. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 139–153). New York: Guilford Press.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Andrew, J. M. (1993). The conceptual basis of antonymy and synonymy in adjectives. *Journal of Memory and Language*, *32*, 301–319.

Nesselroade, J. R., & Molenaar, P. C. M. (1999). Pooling lagged covariance structures based on short, multivariate time-series for dynamic factor analysis. In

R. Hoyle (Ed.), *Research strategies for small samples* (pp. 223–250). Thousand Oaks, CA: Sage.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Read, S. J. (1987). Constructing causal scenarios: A knowledge structure approach to causal reasoning. *Journal of Personality and Social Psychology, 52*, 288–302.

Read, S. J., Jones, D. K., & Miller, L. C. (1990). Traits as goal-based categories: The role of goals in the coherence of dispositional categories. *Journal of Personality and Social Psychology, 58*, 1048–1061.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86*, 61–79.

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology, 83*, 789–803.

Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science, 27*, 767–780.

Rips, L. J., & Estin, P. A. (1998). Components of objects and events. *Journal of Memory and Language, 39*, 309–330.

Skowronski, J. J., Carlston, D. E., Mae, L., & Crawford, M. T. (1998). Spontaneous trait transference: Communicators take on the qualities they describe in others. *Journal of Personality and Social Psychology, 74*, 837–848.

Stapel, D. A., & Koomen, W. (2001). Let's not forget the past when we go to the future: On our knowledge of knowledge accessibility. In G. Moskowitz (Ed.), *Cognitive social psychology: The Princeton Symposium on the Legacy and Future of Social Cognition* (pp. 229–246). Mahwah, NJ: Erlbaum.

Thagard, P., & Kunda, Z. (1998). Making sense of people: Coherence mechanisms. In S. J. Read & L. C. Miller (Eds.). *Connectionist models of social reasoning and social behavior* (pp. 3–26). Mahwah, NJ: Erlbaum.

Todorov, A., Gonzalez, C. M., Uleman, J. S., & Thaden, E. P. (2004). *A dissociation between spontaneous and intentional stereotyped trait inferences*. Manuscript in preparation, New York University.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors: Evidence from false recognition. *Journal of Personality and Social Psychology, 83*, 1051–1065.

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

Trope, Y., & Gaunt, R. (1999). A dual-process model of overconfident attributional inferences. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 161–178). New York: Guilford Press.

Uleman, J. S. (1999). Spontaneous versus intentional inferences in impression formation. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 141–160). New York: Guilford Press.

Uleman, J. S., & Bargh, J. A. (Eds.). (1989). *Unintended thought*. New York: Guilford Press.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wright, J. C., & Mischel, W. (1987). A conditional analysis of dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology*, *53*, 1159–1177.

Wright, J. C., & Mischel, W. (1988). Conditional hedges and the intuitive psychology of traits. *Journal of Personality and Social Psychology*, *55*, 454–469.

Wyer, R. S., Jr., & Lambert, A. J. (1994). The role of trait constructs in person perception: An historical perspective. In P. G. Devine, D. L. Hamilton, & T. M. Ostrom (Eds.). *Social cognition: Impact on social psychology* (pp. 109–142). San Diego, CA: Academic Press.

Zárate, M. A., Uleman, J. S., & Voils, C. I. (2001). Effects of culture and processing goals on the activation and binding of trait concepts. *Social Cognition* [Special issue on culture and cognition], *19*, 295–323.

This page intentionally left blank

PART V

Limits of Mindreading

This page intentionally left blank

17

Mindreading in an Exotic Case

The Normal Adult Human

DALE J. BARR
BOAZ KEYSAR

The world that is most important to humans is the social world—not the mute world of objects, but the world of living, acting minds. Dealing with other minds introduces a fundamental element of uncertainty into life, because the beliefs and desires that drive other people's behavior are hidden from view. Yet, adult humans are highly competent navigators of the social world, because a lifetime of experience with other people has made them experts at reasoning in mentalistic terms. Thus, normal adults can be said to possess a theory of how other minds work that enables them to impute motivations, detect deceptions, and more generally predict and explain the behavior of others. This so-called theory of mind serves as an important foundation for human interaction, making it possible for individuals to coordinate activities, including activities as complex as holding a conversation. An active goal of research in cognitive science is to understand the nature of human "mindreading"—not, of course, mindreading in the magical or paranormal sense but, rather, in the mundane sense of how people apply theory of mind in order to infer the mental states of other people.

Much of what we know about mindreading in humans derives from research on cognitive development. The development of mindreading abilities throughout the lifespan can be characterized as a trajectory away from egocentrism and toward greater and more nuanced mental attribution. Although most normal functioning adults take it for granted that different people can have very different perceptions of reality, developmental research suggests that this understanding unfolds gradually. Many of the rudiments of mindreading seem to be in place well before the third year, such as the ability to discern goal-directed actions (Woodward, 1998) and to appreciate that the behaviors of others is functionally organized in terms of desires (Wellman, 1991). However, an adult-like understanding of other minds is not in place until children appreciate that others can have false beliefs about the world. This critical development is believed to take place between the ages of 4 and 6 years. Children younger than this age seem to have difficulty distinguishing what they believe from what others believe (Wimmer & Perner, 1983), a finding that has been replicated in different cultures with many different tasks (Wellman, Cross, & Watson, 2001).

Another important source of insight into human mindreading abilities comes from research on non-normative and animal cases. The underlying nature of human mindreading abilities has been compared and contrasted with those of chimpanzees (Premack & Woodruff, 1978). Currently, controversy surrounds the question of whether these evolutionarily close relatives have a human-like theory of mind (Povinelli & Vonk, 2003; Tomasello, Call, & Hare, 2003). Researchers have also explored the mindreading and communicative abilities of brain-damaged individuals and have proposed theories regarding the neural circuits that underlie these abilities in normal populations (e.g., Sabbagh, 2004).

In much theorizing on mindreading in humans, the normal adult human appears largely as a figure in the background, an ideal against which the abilities of the chimpanzee, the child, and brain-damaged individual are compared. However, much of what is known about mindreading processes in normal human adults is indirect and based on extrapolation from disordered or child populations—not to mention a lot of common sense. In this chapter, we address the question: What can be learned from directly studying theory-of-mind processes in adults?

To provide an overview, we discuss findings from a research program investigating an important aspect of adult mindreading—a listener's reasoning about the beliefs of a speaker in order to resolve ambiguity in conversation. This research program has revealed some unexpected limits on this aspect of adult mindreading. In particular, our investigations indicate a large degree of egocentrism in how normal adult listeners interpret a speaker's utterances. Across a range of experiments and tasks, adult listeners do not appear to reliably take into account the knowledge that they share with the speaker when they interpret the meaning of what that speaker has said. Given the assumption that communication is typically successful, this seems to indicate that mindreading might play less of a role in certain aspects of language processing than is generally assumed. We suggest that reasoning about a speaker's beliefs occurs primarily as part of self-monitoring, a process through which listeners monitor their interpretations for errors and correct them when they are detected.

Our findings also suggest that the egocentrism observed in 3-year-old children and in individuals with frontal brain damage is also present in normal adults. By using eyetracking techniques, we have been able to observe the process of interpreting another's actions as it unfolds through time. This online methodology suggests that the egocentrism of the young child does not disappear, but lives on in the early moments of adult processing. Although the end product of adult processing shows more sensitivity to others' beliefs, this is not because adults are less likely than children to initially interpret a speaker's action egocentrically. Rather, it is because they are more likely to detect and correct interpretive errors before they *act* upon them. Thus, our data suggest an underappreciated continuity in social reasoning between young children and normal adults.

For the sake of clarity, we begin by noting that the term "mindreading" has been used rather liberally to denote a broad variety of activities related to social cognition. As Ames (2004; Chapter 10, this volume) makes clear, mindreading can involve the application of abstract schematic knowledge or stereotypes, the projection of one's own beliefs or desires onto a target individual, or active perspective taking. In addition, there are marginal cases such as automatic mimicry and emotional contagion (see Hodges & Wegner, 1997, for a review), which may or may not be considered a form of mindreading. One factor that varies across these activities is the extent to which they require the use of metarepresentations, that is, representations about another person's representations. In this chapter, what we are interested in is how people use metarepresentations—specifically, their beliefs about another person's beliefs—in interpreting the actions or utterances of that person.

There is an important distinction between the processes underlying the use of metarepresentations versus the application of stereotypic knowledge structures such as schemata or scripts in order to predict another's behavior. When Albert's roommate Brenda sees him get up from his chair in the living room and walk toward the refrigerator, she might recognize this pattern of behavior as an instance of a well-known script, "going to the fridge." Activation of this script might lead her to impute to Albert the goal of getting something to eat or drink. Although the end result might be a metarepresentation about what Albert desires, the process by which this inference was drawn may not have required metarepresentation. As work by Read and Miller (1998) has shown, this form of inference can be approximated by a neural network. The architecture of their network is based upon the interactive activation network that McClelland and Rumelhart (1981) used to simulate the perception of written words. Thus, the mechanisms underlying schema- or script-based inferences about goals might be identical to the domain-general pattern recognition mechanisms involved in perception. However, if Brenda has reason to believe that Albert thinks the fridge is empty, then the task of imputing a motive to Albert shifts from one of pattern recognition to one of problem solving and decision making. Under such circumstances, Brenda must try to see the world through Albert's eyes in order to predict and explain his behavior, because there is no prior pattern to which the behavior can be matched. She must attempt to retrieve relevant knowledge about Albert, such as the fact that he wants to go downtown, and is walking toward the fridge in order to consult the bus schedule that is attached to its door. In contrast to the pattern recognition case, Brenda's interpretation of Albert's behavior in this situation would seem to involve particularized inferences about Albert's current state of mind. It is this latter sense of mindreading that we address in this chapter.

Studies of how people draw and use metarepresentational inferences of this sort have typically focused on children. However, the normal adult has not been entirely neglected as an object of research. For example, researchers have addressed the neural substrates of theory of mind using normal adults (for a review, see Siegal & Varley, 2002). Another area in which adult perspective taking has been investigated is social attribution (for a review, see Nickerson, 1999). This research has been focused on questions such as how people estimate how they are seen by others (Gilovich, Savitsky, & Medvec, 1998), how people reason about others' construal of a prevailing situation (Griffin, Dunning, & Ross, 1990), or how people impute knowledge to people in a particular social group (Fussell & Krauss, 1991). Much of this work suggests that the outcome of the social reasoning processes is egocentrically biased, with people assuming that other people know the things that they know (Nickerson, 1999).

Although these studies continue to provide valuable insights into social judgment, the cognitive mechanisms by which people put such judgments to use have not been fully specified. What is lacking is a detailed processing model of theory of mind and an understanding of how this system interfaces with other cognitive systems, such as the language processing system. Experimental studies with normal adults can play a critical role in the development of such a model, because they allow for clear isolation of the individual factors that might be involved in mindreading.

Our own interest in the metarepresentational aspect of mindreading derives from our research on how people resolve ambiguity in language comprehension. Theories of language use have long placed metarepresentations at the heart of communication (Clark & Marshall, 1981; Grice, 1957). A basic tenet of modern theories of language use is that language is inherently ambiguous—the same utterance can mean different things, depending upon the speaker's intention. Speakers and listeners can reduce this ambiguity by processing utterances against the background of mutual knowledge, or *common ground*—the set of information that is shared, and critically, *known to be shared* (Clark & Carlson, 1981; Clark & Marshall, 1981). This emphasis on shared knowledge implies that language users maintain and routinely consult particularized models of what their interlocutors know when they process language. Indeed, Clark and Marshall (1981) cogently argue that processing utterances against common ground is the only true guarantee that a communicative act will succeed. Given the intuition that language users are routinely successful at achieving shared understanding, one might expect to find that metarepresentations strongly constrain how normal adults process language.

We began our research by asking the following question: How does a listener's knowledge about what the speaker knows affect how the listener interprets what the speaker says? To answer this question, we have used eye-tracking techniques to monitor the listener's comprehension process (e.g., Keysar, Barr, Balin, & Brauner, 2000). In these studies, participants played the role of "listener" in a communication game with a confederate, who played the role of "speaker." The speaker and listener worked together to rearrange a set of objects in a mutually visible, vertical set of shelves, or "grid," that the experimenter placed between them (see Figure 17.1). The speaker received a diagram showing how the objects should be positioned in this grid. The task was for the listener to follow the speaker's instructions to move objects from slot to slot in order to match the diagram. We occluded the contents of certain slots in the grid so only the listener, but not the speaker, could view them. This created a difference in perspective between the interlocutors, such that only a subset of the objects that the listener knew about was also known to the speaker. For instance, one of our grids contained a large and a medium-sized candle that were mutually visible to both the speaker and the listener. This same grid also had a slot that was occluded from the speaker's view, which contained an even smaller candle that was visible only to the listener. At a certain point the speaker delivered an instruc-

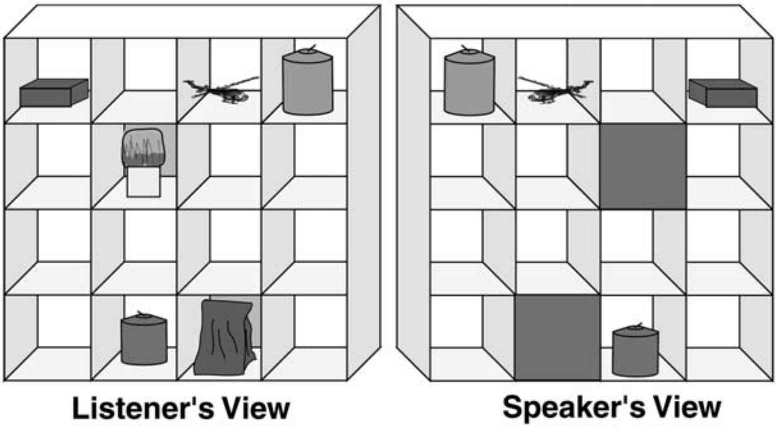

FIGURE 17.1. A grid used in the study from the listener and the speaker's view. The competitor (small candle) or control object is hidden in the bag in the bottom row.

tion that we call the "critical instruction." For example, the speaker instructed the listener to move the "small candle." Note that from the listener's perspective the best match for this expression is the smallest of the three candles. However, the listener does not have any reason to believe that the speaker knows of the existence of this smallest candle. Therefore, he or she should identify the medium-sized candle as the one intended by the speaker (henceforth the "target" object), because that is the smaller of the two candles that the speaker knows about. The listener should ignore this smaller, hidden candle (the "competitor"), even though it is a better match to the speaker's expression, because it is not in common ground. To assess the extent to which listeners still considered the competitor object, we contrasted this test condition with a control condition in which the competitor was replaced with an object that did not match the speaker's expression (e.g., a toy monkey). Our question was how effectively listeners could make use of their common ground with the speaker in order to identify the target.

Our initial studies using this paradigm revealed that listeners strongly considered competitors as referents for the speaker's expressions. For example, they spent more time looking at the competitor (small candle) than the control object (toy monkey), even though they knew the speaker was ignorant of the identity of hidden objects (see Keysar et al., 2000, for details; Keysar & Barr, 2002, for a review of related findings). Occasionally (in about 20% of cases), listeners even attempted to pick up and move competitor objects, although they tended to eventually select the target. These findings suggest a surprising amount of egocentrism in how normal adults interpret referential expressions.

However, do these findings reflect a genuine failure to consider the speaker's perspective, or is there some more compelling alternative explanation? We conducted a set of studies intended to rule out certain alternative explanations and test the robustness of this egocentrism (for a full report, see Keysar, Lin, & Barr, 2003). For example, one such explanation is that perhaps listeners are not egocentric per se, but have difficulty ignoring competitors that are, so to speak, staring them right in the face. More commonly in real-world communication, the non-common ground "competitors" include things that the listener knows about but that are not perceptually available at the time of speaking. Thus, in the experiment described below we sought to eliminate the perceptual availability of competitors by having listeners hide them not only from the speaker but also *from themselves* by placing these objects inside a brown paper bag. The speaker could hear the rustling of the bag, but because of a visual barrier she could not witness the listener hiding the object. Thus, the listener would know of the existence of the competitor, but would not be able to see it at the moment of the critical instruction. If the listener still considered the competitor, this would suggest that the effect is not due to some sort of low-level perceptual interference, but is due to the listener's egocentrism.

We also sought to sharpen the listener's awareness that the speaker's perspective was different by including a condition in which listeners were led to believe that the speaker was not ignorant of the true identity of the occluded object but, instead, had a false belief about it. For example, after the experimenter gave the listener a small candle to hide in the bag, she ostensibly misled the speaker regarding its identity by showing her a picture of a different object (e.g., a small plastic truck) and telling her that this was the object that the listener was hiding. The listener saw the picture and witnessed the experimenter showing it to the confederate. Thus, the listener would be led to believe that the speaker thought that the object in the bag was a truck when it really was a small candle. Listeners were in collusion with the experimenter, who had secretly instructed them beforehand that she would occasionally mislead the speaker about the identity of the object in the bag, and requested that they not reveal its true identity. We contrasted this *false-belief* condition with a condition similar to our previous studies, in which the speaker was not provided any information about the identity of the hidden object (*ignorance condition*).

Would listeners still consider competitors as referents for the speaker's expressions even though they were no longer perceptually available? Furthermore, would they still do so even when they believed that the speaker had a false belief about the identity of the hidden object? We assessed the consideration of hidden objects by examining various eye-tracking measures from the onset of the critical phrase (e.g., "the small candle") up to the moment at which listeners selected the target object. As in previous studies (e.g., Keysar et al., 2000), the pattern of results revealed a strong degree of egocentrism in our normal adult participants. Even though the competitors were not perceptually available (e.g., the candle was in a bag), they still caused substantial interference. When the occluded slot contained a competitor (e.g., the smallest of the three candles), listeners fixated the occluded slot four times as often and for over three times as long as they did when it contained a control object (e.g., the monkey). They also took significantly longer to notice the intended referent (e.g., the smaller of the two larger candles) when the bag contained a competitor.

Most surprisingly, in about 25% of the trials in which the bag contained a competitor, listeners behaved completely egocentrically, picking up the bag and moving it instead of the target object, and had to be corrected by the confederate. Viewed more optimistically, that means that 75% of the time, listeners spontaneously chose the correct target object. However, such optimism must be tempered by the fact that a full 71% of participants moved a competitor on at least one out of the four trials containing a competitor. After listeners realized their blunder, they became much less likely to do it a second time. Moreover, it is important to consider not only the end product of interpretation but also the process that yielded that interpretation. Even though 71% of our listeners made at least one error, nearly all (95%) were delayed in selecting the target when there was a competitor, regardless of whether they moved it or even looked at it. Thus, evidence for egocentrism was present in the eye movement data even when it would not have been revealed by the listener's final judgment.

Even more surprising was that listeners appeared to experience just as much interference from the competitor when they thought that the speaker was *misinformed* about the identity of the competitor as when they thought he or she was ignorant. In other words, even when the listener thought that the speaker believed that the hidden small candle was really a toy truck, upon receiving the instruction to "move the small candle," they looked at the hidden bag containing the candle just as much as they did when they thought that the speaker was ignorant about the contents of the bag. They were also just as likely to move the competitor in the false-belief condition as in the ignorance condition. Relative to the control condition, they were also equally slow to realize that the smaller of the two mutually available candles was the only plausible referent from the speaker's perspective.

These findings highlight the point that even when listeners are clear about what the speaker believes, they have trouble putting this information to use in interpreting the meanings of their actions. To account for such effects, we have put forth the perspective adjustment model (Keysar & Barr, 2002; Keysar, Barr, & Horton, 1998). Perspective adjustment is an anchoring and adjustment model in the spirit of the decision model of Tversky and Kahneman (1974). According to perspective adjustment, listeners initially "anchor" their interpretations of expressions in available information without regard to the other's perspective. By using common ground, listeners can adjust toward the speaker's perspective, although this adjustment is optional—and typically insufficient. This means that there is a systematic source of misunderstanding in language comprehension—namely, a listener's failure to sufficiently discount information not known to the speaker. (For a related model, see Epley, Keysar, Van Boven, & Gilovich, 2004.) As the structure of this model makes clear, we believe that people do make mentalistic inferences during conversation; however, such inferences are not automatic or obligatory, and may have a limited role in guiding the earliest moments of comprehension.

The possibility that inferences about common ground are optional, or are made largely as a kind of afterthought, means that there is no guarantee that communication will be successful in any given case. However, this might not be such a bad thing if what is typically available to the listener also tends to be available to the speaker. If this is the case, then access to a speaker's beliefs may not be a necessary prerequisite to successful communication. Such an idea is admittedly controversial, given the avowed ambiguity of communication (although see Recanati, 2002, for a philosophical defense of this position). As opposed to the potential ambiguity that is latent in any given utterance, the *actual* ambiguity that language users experience is a function of the degree of alignment between their perspectives. At this point, unfortunately, we know little about the factors that cause perspectives to diverge or converge, because this is still an emerging line of research. Existing research does, however, claim that the perspectives of a speaker and a listener may come into alignment through low-level, resource-free implicit mechanisms such as priming and associative learning (Barr, 1999; Barr & Keysar, 2002; Garrod & Anderson, 1987; Garrod & Doherty, 1994; Markman & Makin, 1998; Pickering & Garrod, 2004). Although processes such as associative learning or priming might seem too simplistic to hold things together during a conversation, recent research using multiagent computer simulation indicates that such mechanisms may be sufficient to support coordinated communication in a community where mindreading is absent (Barr, 2004; Steels, 1998).

Furthermore, the egocentrism of language users makes sense when assessed against the feedback-rich environment of situated language use (Barr & Keysar, 2004). Because language users consistently monitor and provide feedback to their interlocutors (Clark & Brennan, 1991; Clark & Krych, 2004), there are many opportunities in natural dialogue to interactively diagnose and correct miscoordination. Of course, how people actually use this feedback to improve coordination is likely to involve access to common ground, which is precisely the role that the perspective adjustment model stipulates for this kind of knowledge. We hope that this observation drives home the point that our argument is not about whether people ever draw metarepresentational inferences during conversation—clearly, they do—but concerns *when* they do and *how* they put these inferences to use.

Our findings, which suggest that mental state attribution may not be fully integrated into language processing, are consistent with the observation that children become sophisticated language users before they become sophisticated mindreaders. Thus, a child 4 years old who would fail the false-belief task would have little trouble understanding and producing references in speech, even though the metarepresentational capacity has been viewed as a *sine qua non* of successful reference (e.g., Clark & Marshall, 1981). In fact, Epley, Morewedge, and Keysar (2004) demonstrate that adults are different from children not in the initial egocentric process but in their ability to effectively recover from an error. Using a paradigm similar to the one we described, Epley and colleagues show that adults look at the hidden competitor just as quickly as children, but they are faster to identify the target and less likely to move the competitor. This demonstrates that the early moments of comprehension are the same for children and adults. However, a critical difference between the adult and child is the adult's ability to self-monitor and preempt or recover from an egocentric error.

The study of mindreading in normal adults suggests a continuity in processing not only between children and adults but perhaps even between the normal case and people with prefrontal brain damage. Research on such patients suggests that these individuals have difficulty in inhibiting prepotent responses that are cued by environmental stimuli but irrelevant to their current goals (see Miller & Cohen, 2001, for a review). For example, individuals with prefrontal damage produce more errors on the Stroop task than normals (Vendrell et al., 1995). We have found a task that, in essence, can cause normal individuals to occasionally behave in ways that are similar to such patients, in the sense that listeners have difficulty inhibiting the selection of the competitor, which was always a better referent for the speaker's expression. An interesting prediction from our study is that if adults are placed under a severe cognitive load that would inhibit their ability to self-monitor, the difference between normal adults and frontal patients might be diminished. The load manipulation might have little effect, however, on the speed of performance in the control condition, suggesting that the core language abilities would remain intact. Finally, it is an interesting question whether adults under cognitive load might perform like children on false-belief tasks or other tasks involving mental attribution. This result might be expected, considering that cognitive load has been shown to interfere with perspective-taking processes (Hodges & Wegner, 1997).

To conclude, we hope to have presented some compelling arguments why adult theory of mind should be an object of investigation in its own right. Research has only scratched the surface of the complexities of the cognitive mechanisms underlying mindreading in the normal adult. Adults represent the endpoint of development, and as such they provide a context for understanding developments in the young child. Likewise, to understand what mechanisms are absent or impaired in the case of people with mindreading deficits, such as in the case of prefrontal damage, it is important to have a standard of comparison that is empirically grounded. Currently, much research tacitly assumes an adult mindreading competency that is sophisticated, routinely accessed, and tightly integrated with other cognitive functions. Against the background of such an ideal, the egocentric behavior of the normal adults that we have observed in our laboratory appears quite exotic indeed.

REFERENCES

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

Barr, D. J. (1999). *A theory of dynamic coordination for conversational interaction.* Unpublished doctoral thesis, University of Chicago.

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, *28*, 937–962.

Barr, D. J., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language, 46*, 391–418.

Barr, D. J., & Keysar, B. (2004). Making sense of how we make sense: The paradox of egocentrism in language use. In H. L. Colston & A. N. Katz (Eds.), *Figurative language comprehension: Social and cultural influences*. Mahwaw, NJ: Erlbaum.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Clark, H. H., & Carlson, T. B. (1981). Context for comprehension. In J. Long & A. Baddeley (Eds.), *Attention and performance* (Vol. 9, pp. 313–330). Hillsdale, NJ: Erlbaum.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*, 62–81.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshe, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge, UK: Cambridge University Press.

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology, 40,* 760–768.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*, 327–339.

Fussell, S. R., & Krauss, R. M. (1991). Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology, 21*, 445–454.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*, 181–218.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition, 53*, 181–215.

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology, 75*, 332–346.

Grice, H. P. (1957). Meaning. *Philosophical Review, 66*, 377–388.

Griffin, D. W., Dunning, D., & Ross, L. (1990). The role of construal processes in overconfident predictions about the self and others. *Journal of Personality and Social Psychology, 59*, 1128–1139.

Hodges, S. D., & Wegner, D. M. (1997). Automatic and controlled empathy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 311–339). New York: Guilford Press.

Keysar, B., & Barr, D. J. (2002). Self-anchoring in conversation: Why language users don't do what they "should." In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 150–166). Cambridge, UK: Cambridge University Press.

Keysar, B., Barr, D. J., Bailin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*, 32–38.

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science, 7*, 46–50.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25–41.

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General, 127*, 331–354.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375–405.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience, 24*, 167–202.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*(6), 737–759.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169–226.

Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: Suspiciously human? *Trends in Cognitive Sciences, 7*, 157–160.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*, 515–526.

Read, S. J., & Miller, L. C. (1998). On the dynamic construction of meaning: An interactive activation and competition model of social perception. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and social behavior* (pp. 27–70). Mahwah, NJ: Erlbaum.

Recanati, F. (2002). Does linguistic communication rest on inference? *Mind and Language, 17*, 105–126.

Sabbagh, M. A. (2004). Understanding orbitofrontal contributions to theory-of-mind reasoning: Implications for autism. *Brain and Cognition, 55,* 209–219.

Siegal, M., & Varley, R. (2002). Neural systems involved in "theory of mind." *Nature Reviews Neuroscience, 3*, 463–471.

Steels, L. (1998). Synthesizing the origins of language and meaning using coevolution, self-organization and level formation. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language: Social and cognitive bases* (pp. 384–404). Cambridge, UK: Cambridge University Press.

Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees understand psychological states—the question is which ones and to what extent. *Trends in Cognitive Sciences, 7*, 153–156.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Vendrell, P., Junque, C., Pujol, J., Jurado, M. A., Molet, J., & Grafman, J. (1995). The role of prefrontal regions in the Stroop task. *Neuropsychologia, 33*, 341–352.

Wellman, H. M. (1991). From desires to beliefs: Acquisition of a theory of mind. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading* (pp. 19–38). Oxford, UK: Blackwell.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*, 655–684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*, 1–34.

18

Empathy Gaps in Emotional Perspective Taking

LEAF VAN BOVEN
GEORGE LOEWENSTEIN

One of the hallmarks of psychological maturity is a "theory of mind" that allows one to understand that other people, particularly those in different situations, have feelings, preferences, and behavioral inclinations that are different from one's own. Having just gorged on Ritz crackers and Velveeta cheese, one expects that someone who has not eaten in several hours is hungrier than oneself. And even though one may have just received good news that a manuscript was accepted, there is no surprise when someone whose grant was not funded is not elated about one's success. In each of these situations, a theory of mind creates the recognition and anticipation that, because others are in a different affective situation, they feel, think, and behave differently than oneself.

Such "emotional perspective taking" is ubiquitous in everyday life, and doing it well is important for social interactions. This chapter describes a simple, dual-judgment model of how—and how well—people engage in emotional perspective taking. The chapter also describes recent studies that test a key implication of the model: that errors and biases in predicting one's own reactions to emotional situations produce corresponding errors and biases in predicting others' reactions to emotional situations.

DUAL JUDGMENTS IN EMOTIONAL PERSPECTIVE TAKING

Social psychologists are fond of pointing out that, despite their sophisticated theory of mind, adults are more egocentric than they like to believe, and that social judgments tend to be biased in the direction of their own attitudes, behaviors, and beliefs (e.g., Krueger & Clement, 1997; Marks & Miller, 1987; Nickerson, 1999, 2001; Royzman, Cassidy, & Baron, 2003). Most social-psychological research on adult perspective taking has focused on people's judgments about other people who are in a similar, nonemotional situation as the self—for example, when others are faced with a similar decision (L. Ross, Greene, & House, 1977), given a similar personality test (Krueger & Clement, 1994), or asked a similar question about their cinematic preference (Gilovich, 1990). This research, represented by the horizontal dashed arrow in Figure 18.1, has generally shown that people tend to overestimate how similar other people are to themselves (Krueger & Clement, 1994). Much previous research has therefore sought to understand the causes of this overestimation.

Some research suggests, for example, that people overestimate the similarity between themselves and others because of the often reasonable assumption that the self is informative about others (Dawes, 1989, 1990; Hoch, 1987). People may also overestimate the similarity between themselves and others because they tend to selectively associate with similar others (Bosveld, Koomen, & van der Pligt, 1994; Sherman,

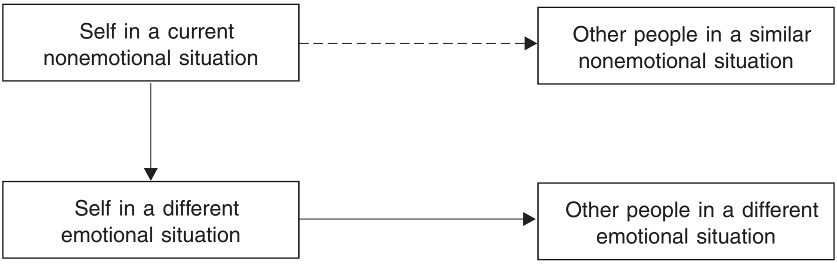
Self in a current nonemotional situation
Other people in a similar nonemotional situation
Self in a different emotional situation
Other people in a different emotional situation

FIGURE 18.1. Graphical representation of perspective-taking process when other people are in a similar nonemotional situation (the rightward-pointing dashed arrow) and when other people are in a different emotional situation (the downward- and rightward-pointing solid arrows).

Chassin, Presson, & Agostinelli, 1984). Or people may inflate others' similarity to themselves out of a desire to be "normal" or to otherwise maintain favorable self views (Krueger & Clement, 1994; L. Ross et al., 1977). Finally, people may overestimate the similarity between themselves and others because they believe that they respond to the objective properties of the situations they encounter rather than to their subjective constructions of those situations, and assume, therefore, that others will respond similarly (Gilovich, 1990; Gilovich, Jennings, & Jennings, 1983; Griffin & Ross, 1991; L. Ross & Ward, 1995).

As noted, most previous research examines people's predictions of the behavior of others who are in similar, typically nonemotional, situations as themselves. By focusing on judgments about others who are in similar situations, previous research is ill equipped to explain emotional perspective taking, which entails predicting the reaction of people who are in emotional situations different from the situation the self is currently in (as indicated by the lower-right box in Figure 18.1).

We propose that such emotional perspective taking entails two distinct judgments, represented by the two solid arrows in Figure 18.1 (Van Boven & Loewenstein, 2003; Van Boven, Loewenstein, & Dunning, 2004, 2005). The first is a prediction of what one's own preferences and decisions would be in the other person's emotional situation (the downward-pointing vertical solid arrow in Figure 18.1). The second judgment is an assessment of how similar the other person is to the self, and, hence, how informative self-predictions are about the other person's reactions (the horizontal solid arrow in Figure 18.1).

According to this dual judgment model, the accuracy of emotional perspective taking depends importantly on both judgments. Previous research has amply documented that people tend to overestimate how similar they are to other people (Krueger & Clement, 1997; Marks & Miller, 1987) so that, even if they were perfectly calibrated in predicting their own reactions to emotional situations, people would probably make biased predictions about others' preferences and decisions. Notice, however, that even if people were perfectly calibrated in judging how similar they are to other people, the accuracy of their social predictions also depends on the accuracy of their self-predictions. Because people tend to view others as similar to themselves, if they mispredicted their own reactions to emotional situations, these biased self-predictions will produce biased social predictions.

We have recently tested two hypotheses derived form this dual-judgment model of emotional perspective taking. The first is that people should explicitly report using self-predictions as a basis for predicting others' reactions to emotional situations. The second is that people make biased judgments of their own reactions to emotional situations, and that these biased self-predictions lead to biased predictions of others' reactions to emotional situations. Specifically, because people in "cold," nonemotional states underestimate the impact of "hot," emotional arousal on their own preferences and decisions, people in cold states should also underestimate the impact of being in a hot state on others' preferences and decisions.

SELF-PREDICTION AS SOCIAL PREDICTION

Several studies demonstrate that people explicitly report using self-predictions as a basis for emotional perspective taking. In one, people were shown a picture of three hikers (the two authors and Douglas Harsch) trudging through an Alaskan mountain meadow (Van Boven & Loewenstein, 2003, Study 1). Participants read a scenario describing how, through an unfortunate encounter with a bear, a 12-day backpacking trip through the Alaskan wilderness went bad and the three hikers were forced to spend 4 days without food, navigating their way back to civilization—a trek that involved crossing a large glacier, traversing a raging river, and bushwhacking through a dense forest.

After a few minutes spent thinking about what the hikers felt during their ordeal, participants described in their own words "the processes and strategies you used to imagine what the hikers were thinking and feeling." Consistent with our model, most people (79%) explicitly described mentally trading places with the hikers, predicting how they would react to being in the hikers' situation. Furthermore, most people (69%) mentioned something about their own reactions to the hikers' situation ("I would be really scared and hungry . . . ") before they mentioned something about the hikers' reaction ("so I bet the hikers were really scared and hungry"). In other words, when asked "How would the hikers feel?" people answered with "If I were the hikers, I would feel . . . ."

In many cases, of course, self-predictions are a good source of evidence when making judgments about other people (Dawes, 1989, 1990; Hoch, 1987). When little is known about others' reactions to an emotional situation, one's own anticipated responses are reasonable proxies for others' responses. In other cases, people may have sophisticated theories that guide their predictions of others they know well (e.g., "Doug seems unaffected by outdoor risks, so the bear probably wouldn't faze him"). Or people may rely on intuitions about how people generally respond to specific situations ("Any reasonable person would be terrified of a face-to-face confrontation with a bear, so the hikers were probably scared witless"). We suspect, however, that people are strongly inclined to use themselves as a basis for making predictions about others even when they have evidence that their own reactions are anomalous (Krueger & Clement, 1994) and even when they should recognize that their own experiences are of limited relevance—for example, when others' experience is blatantly different from their own.

To illustrate this point, consider the following simple study (Van Boven, Loewenstein, & Dunning, 2004). Participants read about one of two protagonists. One was Tom, who experienced a near-death bout with testicular cancer. The other was Shelia, a young woman whose son nearly died during a difficult childbirth. Participants were asked to spend a few minutes thinking about the protagonist's thoughts and feelings. We expected participants' gender to moderate the degree that they viewed themselves as informative about the protagonist. Because females cannot (without major surgery) become males, and hence cannot suffer testicular cancer, we expected women to judge themselves as less informative about Tom's experience than males. And because males cannot (without major surgery) become females and give birth, we expected men to judge themselves to be less informative about Shelia's experience than women. Indeed, males reported mentally trading places with Tom (87%) more than with Shelia (74%); and females reported mentally trading places with Shelia (87%) more than with Tom (75%). These differences are rather small, however. In fact, the majority of people reported that they mentally traded places with the protagonist even when it was impossible for them ever to experience exactly what the protagonist experienced: most males traded places with Shelia, and most females traded places with Tom.

These results suggest that people are inclined to use themselves as a basis for predicting others' reactions to emotional situations even when their own experiences are not—without a large stretch of the imagination—relevant. The intuitive appeal of the self as a basis for judging others, we suspect, stems largely from the fact that people have well-developed and highly accessible, albeit often incorrect, theories of and knowledge about the self.

EMPATHY GAPS IN SELF AND SOCIAL PREDICTIONS

If people use their self-predictions as a basis for making predictions about others' reactions to emotional situations, then the accuracy of social predictions should depend critically on the accuracy of self-predictions. Previous research indicates that predictions of one's own reactions to emotional situations are systematically biased. In particular, people in a cold, emotionally unaroused state often have difficulty bridging the gap between their current preferences and decisions and what their preferences and decisions would be in a hot, emotionally aroused state (Loewenstein, 1996; Loewenstein, O'Donoghue, & Rabin, 2003). For example, men who are not sexually aroused predict they would be less likely to engage in sexually aggressive behavior than men who are sexually aroused (Loewenstein, Nagin, & Paternoster, 1997). People who are sated because they have just eaten are less likely than hungry people who have not eaten to choose a high-calorie snack to consume just after a future lunch (Read & van Leeuwen, 1998). And people who are hungry because they have not eaten expect to be more interested in eating a plate of spaghetti for breakfast than people who are sated (Gilbert, Gill, & Wilson, 2002). These *empathy gaps* occur largely because people believe that their preferences and decisions are based primarily on the inherent desirability of choice alternatives rather than affectively influenced constructions of those alternatives (Griffin & Ross, 1991; Pronin, Gilovich, & Ross, 2004; L. Ross & Ward, 1996).

Our dual-judgment model of emotional perspective taking implies that these empathy gaps in self-predictions produce corresponding empathy gaps in social predictions. We have recently tested this prediction in studies that involved bodily drive states and self-conscious emotions.

In one study, we asked people to read a description of three hikers lost in the Colorado mountains without food or water, to predict whether hunger or thirst would be more distressing to the hikers, and to predict which the hikers would regret more, not bringing water or not bringing food on their hike (Van Boven & Loewenstein, 2003). We also asked people to predict what would be more distressing to themselves and what they would regret more if they were in the hikers' situation. We randomly assigned people entering an exercise facility to answer these questions either immediately before or immediately after engaging in vigorous cardiovascular activity for at least 20 minutes, which we assumed would make them thirsty and warm.

Consistent with our model of emotional perspective taking, people's exercise-induced thirst influenced their predictions of how the hikers (and they themselves) would feel. Nearly everyone who had just exercised expected that both they themselves and the hikers would be more bothered by thirst and would regret not bringing water more than not bringing food. In contrast, only about half of the people who were just about to exercise predicted that they themselves and the hikers would be more bothered by thirst and would regret more not bringing water. Although we could not measure hikers' actual feelings, given that people die of dehydration more quickly than starvation, we strongly suspect that thirst would be more bothersome to the hikers than hunger (a suspicion supported by various informal and unintended personal investigations by the two authors).

People who were less thirsty because they had not exercised thus experienced an empathy gap when making predictions about the hikers that mirrored the empathy gap they experienced in predicting their own preferences. Additional path analyses indicated that, statistically speaking, the empathy gaps in self-predictions fully explained the empathy gaps in people's predictions of the hikers' feelings. After accounting for the impact of exercise on self-predictions, there was no residual impact of exercising on people's predictions of the hikers' feelings.

We have also shown that people experience empathy gaps when predicting how other people would react to situations that arouse self-conscious emotions. In particular, embarrassment and the desire to avoid it is a powerful psychological restraining force: many important failures to act can be attributed to fear of embarrassment, including nonintervention in emergency situations (Latane & Darley, 1970) and nonopposition to unpopular policies or social norms (Miller & McFarland, 1987; Prentice & Miller, 1993; Van Boven, 2000).

Despite the frequency with which people confront embarrassing situations, we have found that they systematically underestimate the impact that fear of embarrassment would have on their preferences and decisions. Specifically, when embarrassing public performances are purely hypothetical or in the psychologically distant future, people overestimate how willing they would be to perform, compared to when the performances are real and immediate (Van Boven, Loewenstein, Dunning, & Welch, 2004). According to our model of emotional perspective taking, these empathy gaps in self-prediction should contribute to underestimating the impact of fear of embarrassment on other people.

In one experiment, we asked half of the students in a large lecture class whether they would be willing to dance for 1 minute in front of the rest of the class to Rick James's 1981 funky song "Super Freak" in exchange for $5 (Van Boven et al., 2005, Experiment 2). The other half of the class was asked simply to imagine that they had been given the option of dancing for $5 and to predict whether they would dance if they were actually given the choice to do so. In addition, both groups of students were asked to predict the decision made by a randomly selected student (other than themselves) who actually faced the choice of dancing for money.

As in our previous research (Van Boven, Loewenstein, Dunning et al., 2004), a larger fraction of students facing a hypothetical performance predicted they would be willing to dance (31%), as compared with the fraction of students who were actually willing to dance (8%). More important for the present purposes, students who themselves faced a purely hypothetical performance predicted that other students would be more willing to dance (30%) than students who themselves faced a real performance (16%).

As with the study of people's predictions of the hikers' feelings, these results indicate that people who themselves faced a hypothetical decision to dance, and were in a relatively cold state, experienced an empathy gap when predicting others' decision to dance. This empathy gap in social prediction mirrored an empathy gap in self-prediction. Furthermore, subsequent analyses indicated that empathy gaps in self-predictions statistically explained the empathy gaps in predictions of others' willingness to engage in an embarrassing public performance. After accounting for the influence of facing a real or hypothetical decision to dance on self-predictions, there was no residual impact of the real or hypothetical nature of the decision on people's predictions of others' decision to dance.

Notice that students who faced a real performance expected that others would be more willing to dance than they were themselves. This difference between predictions of self and others is consistent with previous research indicating that people tend to believe that others are less influenced by fear of embarrassment than they themselves are (McFarland & Miller, 1990; Sabini, Cosmas, Siepmann, & Stein, 1999; Van Boven, 2000). In terms of our model, people's intuitive theory that others are less influenced by fear of embarrassment than the self influences their assessment of the similarity between their self- and social predictions (the horizontal solid arrow in Figure 18.1). The fact that students expected others to be more willing to dance than themselves illustrates their assessment of similarity—or dissimilarity, in this case—between self and others.

Taken together, our studies indicate that people experience empathy gaps in emotional perspective taking that mirror empathy gaps in self-predictions. People who are in a cold, nonemotional state underestimate the impact of being in a hot, emotionally arousing situation on other people's preferences and behaviors, just as they underestimate the impact of being in a hot state on their own preferences and behaviors.

QUESTIONS AND CONSEQUENCES

Predicting Feelings versus Choices

Our results join a large and growing body of research on predicting the psychological impact of emotional situations. Other researchers have found that people overestimate the intensity and duration of their feelings in response to emotional events (Gilbert & Wilson, 2000; Wilson & Gilbert, 2003). This so-called impact bias result might seem inconsistent with our findings that people underestimate the impact of emotional situations on their preferences and choices. This contradiction may be more apparent than real, however. The impact bias occurs when people in a cold state predict how being in an emotional situation would influence their feelings, whereas empathy gaps occur when people in a cold state predict how being in an emotional situation would influence their preferences and decisions. There is an important conceptual distinction between feelings, the phenomenological manifestation of emotional arousal, versus preferences and decisions, the selection of one alternative over other alternatives (Van Boven & Kane, in press). Because feelings, preferences, and decisions are conceptually distinct, there is not necessarily a logical inconsistency between differences in predictions of feelings versus preferences and decisions.

Still, an intriguing possibility is that people simultaneously overestimate the influence of emotional situations on their own and others' feelings while underestimating the influence of emotional situations on their own and others' preferences and decisions. Testing the veracity of this conjecture is an important task for future research. In any event, the complexities of emotional perspective taking would be substantially advanced by understanding the differences between empathy gaps and impact bias in self-predictions.

Top-Down versus Bottom-Up Perspective Taking

Ames (Chapter 10, this volume) distinguishes between perspective taking from the bottom up and from the top down. In bottom-up perspective taking, perceivers use the raw data of physical actions and social behavior to infer others' mental states. In top-down perspective taking, perceivers use higher-level mental constructs such as theories, introspections, and stereotypes to infer others' mental states, with little or no reference to others' concrete actions. Our model is clearly of the top-down variety: People use their self-predictions as a basis for inferring others' preferences and decisions. Of course, participants in our studies were asked to make judgments about hypothetical or anonymous others, so there was little or no opportunity for them to make bottom-up inferences. An important question, then, is to what extent people use self-predictions as a basis for judging the preferences of other real, specific, "live" individuals who are in emotional situations.

This question, ultimately, is empirical. We suspect, however, that it is extremely difficult for people to set aside their self-predictions when making social predictions and that they are reluctant to do so even when provided with ample behavioral evidence about others' preferences. Self-knowledge is often more accessible than social knowledge: we simply have a greater amount of more easily retrievable knowledge about our own personal history, preferences, attitudes, and beliefs. This differential accessibility may directly increase the weight of the self in emotional perspective taking (Ross & Sicoly, 1979; Taylor, 1982; Tversky & Kahneman, 1973).

Behavioral Misinterpretation

Furthermore, when people are in different emotional situations from their interaction partner, they may have difficulty in taking others' behavior at face value, as it were. When other people are in an emotional situation, it may be difficult *not* to predict how oneself would react (Hodges & Wegner, 1997). This self-prediction, once made, is likely to serve as an (erroneous) expectation against which others' behaviors are judged (Reeder, Fletcher, & Furman, 1989). In our study of dancing for money, for example, a student facing a hypothetical choice may use her erroneous prediction that she would dance for $5 as a basis for inferring that another's decision *not* to dance reflects the nondancer's dispositional shyness rather than a normal reaction to an embarrassing situation. Self-predictions may thus lead people in cold states to misinterpret the actions of people who are in emotional situations.

These misinterpretations, in turn, can cause people in nonemotional states to behave differently toward people in emotional states from how they would if they had a true appreciation of the power of emotion to shape preferences and behavior. Nonaddicted policymakers, for example, may misinterpret crimes induced by drug craving as caused by the perpetrators' dispassionate calculations of costs and benefits—a misinterpretation that could foster policies of deterrence and punishment rather than treatment and prevention.

CONCLUSION

Conventional psychological wisdom holds that social judgment is egocentric: judgments of others are made in comparison to the self, in service of the self, and in the direction of the self. Much of the conventional evidence for this wisdom is the strong correlation between judgments of the self and others (Krueger, 1998). The status quo has recently been challenged with the claim that correlations between self and social judgments are the spurious result of prototypes (Karniol, 2003) or implicit theories (Gopnik, 1993) about the way minds work. A central feature of these challenges is that both self- and social judgments are based on a single stable prototype or theory of how people feel, think, and behave in different situations.

The results of our studies present a strong case that social judgments can be truly egocentric. Our results indicate that arousing emotions in oneself influences predictions of others' preferences and decisions—evidence that is difficult to explain from a prototype or implicit theory point of view. Why would exercise affect people's theory or prototype of lost hikers' thirst? Why would facing a real, as opposed to hypothetical, embarrassing performance influence theories or prototypes of whether university students would dance for money? We concur that correlations between self- and social judgments are often misinterpreted as indicators of causal relationships when they may simply reflect judgments based on prototypes, theories, or response biases. But our model and, more importantly, the results of our studies indicate that genuinely egocentric social judgment, at least in the case of emotional perspective taking, is alive and well.

REFERENCES

Bosveld, W., Koomen, W., & van der Pligt, J. (1994). Selective exposure and the false consensus effect: The availability of similar and dissimilar others. *British Journal of Social Psychology, 33*, 457–466.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology, 25*, 1–17.

Dawes, R. M. (1990). The potential nonfalsity of the false consensus effect. In R. M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 179–199). Chicago: University of Chicago Press.

Gilbert, D. T., Gill, M. J., & Wilson, T. D. (2002). The future is now: Temporal correction in affective forecasting. *Organizational Behavior and Human Decision Processes, 88*, 430–444.

Gilbert, D. T., & Wilson, T. D. (2000). Miswanting: Some problems in the forecasting of future affective states. In J. Forgas (Ed.), *Feeling and thinking: The role of affect in social cognition* (pp. 178–198). Cambridge, U.K.: Cambridge University Press.

Gilovich, T. (1990). Differential construal and the false consensus effect. *Journal of Personality and Social Psychology, 59*, 623–634.

Gilovich, T., Jennings, D. L., & Jennings, S. (1983). Causal focus and estimates of consensus: An examination of the false consensus effect. *Journal of Personality and Social Psychology, 45, 550–559.*

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16*, 1–14.

Griffin, D. W., & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 24, pp. 319–359). San Diego, CA: Academic Press.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53*, 221–234.

Hodges, S., & Wegner, D. M. (1997). Automatic and controlled empathy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 311–339). New York: Guilford Press.

Karniol, R. (2003). Egocentrism versus protocentrism: The status of the self in social prediction. *Psychological Review, 110*, 564–580.

Krueger, J. (1998). On the perception of social consensus. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 163–240). San Diego, CA: Academic Press.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67*, 596–610.

Krueger, J., & Clement, R. W. (1997). Estimates of social consensus by majorities and minorities: The case for social projection. *Personality and Social Psychology Review, 1*, 299–313.

Latane, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* Englewood Cliffs, NJ: Prentice-Hall.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65*, 272–292.

Loewenstein, G., Nagin, D., & Paternoster, R. (1997). The effect of sexual arousal on predictions of sexual forcefulness. *Journal of Crime and Delinquency, 32*, 443–473.

Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Projection bias in predicting future utility. *Quarterly Journal of Economics*, 1209–1248.

Marks, G., & Miller, N. (1987). Ten years of research on the false consensus effect: An empirical and theoretical review. *Psychological Bulletin, 102*, 72–90.

McFarland, C., & Miller, D. T. (1990). Judgments of self-other similarity: Just like other people, only more so. *Personality and Social Psychology Bulletin, 16*, 475–484.

Miller, D. T., & McFarland, C. (1987). Pluralistic ignorance: When similarity is interpreted as dissimilarity. *Journal of Personality and Social Psychology, 53*, 298–305.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*, 737–759.

Nickerson, R. S. (2001). The projective way of knowing: A useful heuristic that sometimes misleads. *Current Directions in Psychological Science, 10*, 168–172.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology, 64*, 243–256.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review, 111,* 781–799.

Read, D., & van Leeuwen, B. (1998). Time and desire: The effects of anticipated and experienced hunger and delay to consumption on the choice between healthy and unhealthy snack food. *Organizational Behavior and Human Decision Processes, 76*, 189–205.

Reeder, G. D., Fletcher, G. J., & Furman, K. (1989). The role of observers' expectations in attitude attribution. *Journal of Experimental Social Psychology, 25*, 168–188.

Ross, L., Greene, D., & House, P. (1977). The "false-consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*, 279–301.

Ross, L., & Ward, A. (1995). Psychological barriers to dispute resolution. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 27, pp. 255–304). San Diego, CA: Academic Press.

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *Values and knowledge* (pp. 103–135). Mahwah, NJ: Erlbaum.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology, 37*, 322–336.

Royzman, E. B., Cassidy, K. W., & Baron, J. (2003). "I know, you know": Epistemic egocentrism in children and adults. *Review of General Psychology, 7*, 38–65.

Sabini, J., Cosmas, K., Siepmann, M., & Stein, J. (1999). Underestimates and truly false consensus effects in estimates of embarrassment and other emotions. *Basic and Applied Social Psychology, 21*, 233–241.

Sherman, S. J., Chassin, L., Presson, C. C., & Agostinelli, G. (1984). The role of the evaluation and similarity principles in the false consensus effect. *Journal of Personality and Social Psychology, 47*, 1244–1262.

Taylor, S. E. (1982). The availability bias in social perception and interaction. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment and uncertainty: heuristics and biases* (pp. 190–200). Cambridge, UK: Cambridge University Press.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–232.

Van Boven, L. (2000). Pluralistic ignorance and political correctness: The case of affirmative action. *Political Psychology, 21*, 267–276.

Van Boven, L., & Kane, J. (in press). Predicting feelings and choices. In L. J. Sanna & E. C. Chang (Eds.), *Judgments over time: The interplay of thoughts, feelings, and behaviors*. New York: Oxford University Press.

Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin, 29*, 1159–1168.

Van Boven, L., Loewenstein, G., & Dunning, D. (2004). *Changing places: A theory of empathy gaps in emotional perspective taking*. Unpublished manuscript, University of Colorado, Boulder.

Van Boven, L., Loewenstein, G., & Dunning, D. (2005). The illusion of courage in social-predictions: Underestimating the impact of fear of embarrassment on other people. *Organizational Behavior and Human Decision Processes, 96*, 130–141.

Van Boven, L., Loewenstein, G., Dunning, D., & Welch, N. (2004). *The illusion of courage: Underestimating the impact of fear of embarrassment on the self*. Unpublished manuscript, University of Colorado, Boulder.

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 35, pp. 345–411). San Diego, CA: Academic Press.

19

Is How Much You Understand Me in Your Head or Mine?

SARA D. HODGES

Despite most people's acknowledgment that the subjective border surrounding another person's mind can only be crossed in science fiction, it is a common belief that if two minds have independently experienced similar events, the "mental landscapes" of those two minds may also resemble each other more. This may in turn allow the two minds to know each other better. Perceivers assume they'll understand another person better if they've had similar experiences, saying things like "I've been in your shoes and know just how you feel"; and the targets of their understanding also express wishes that make the same assumption, such as "I wish I could talk to someone who's been through it."

Given the widespread assumption that having had a similar experience will help one person to understand another, embarking on a program of research that aims to investigate this correlation may seem ill advised: the findings would be replications at best, or potentially dismissed as obvious. However, it turns out that there aren't many studies showing that similar experience really does make people more understanding or empathic. Rarer still is work that examines *how* that similar experience helps, if it does indeed make a difference.

When others have investigated this question, the studies are often in the clinical realm, and they examine similarities between clients and therapists that sometimes could be considered “similar experience” but are often better described as shared demographic characteristics. Furthermore, even when limited to the highly specific, ritualized set of interpersonal interactions that occur in the context of therapy, the case for similar experience is very weak. An informal qualitative review of this work suggests that the cases where there is no evidence of experience helping outnumber the studies where experience does seem to help (including some fairly poorly designed ones) by about a 3:1 ratio (see Hodges, Klein, Veach, & Villanueva, 2004, for a review of some of these studies).

In some ways, this lack of empirical evidence makes the question of similar experience all the more intriguing, because it adds another question: If there are few studies showing the advantages of similar experience, then what in our everyday experience leads us to think that someone who has “been there too” will be more understanding? In this chapter, I discuss the results of three studies of empathy and shared experience conducted in my lab, from which a still tentative yet interesting picture about the relationship of similar experience to empathy is starting to emerge.

One thing that seems critical in drawing valid conclusions about the effects of similar experience on empathic understanding is to study the kinds of potent life events for which people believe experience makes a difference. The ideal life events would be important and consequential and, at the same time, the kinds of things that some normal people experience every day (and other equally normal people do not). Thus, the life experiences examined in these studies—new motherhood (using a community sample; Hodges et al., 2004), alcoholism (drawing alcoholics from the community and nonalcoholics from a college student sample; Hallinan, 2000) and parental divorce (using a college student sample; Hodges, 2004)—are not the kinds of experiences that can be experimentally manipulated.

The studies had a number of other similar characteristics and used a largely similar methodology. In all three, the “targets” who had had the life experience (either women who were new to motherhood and had very young firstborn infants, male and female alcoholics who were members of Alcoholics Anonymous [AA], or college students of both sexes with divorced parents) talked about their experience while being videotaped. In the new-motherhood and alcoholism studies, targets answered the open-ended questions of an interviewer who was off camera, whereas in the divorce study targets talked about divorce and its effects on children growing up with another student who either also did or did not have divorced parents.

In all three studies, Ickes’s (2001) empathic accuracy paradigm was used. Targets watched the videotape of themselves and were asked to stop the tape at any point that they remembered having had a thought or feeling. They then recorded the time and the content of the thought or feeling. Additionally, targets in the new-motherhood study completed a set of questions measuring new mothers' adjustment toward their new role—for example, whether they ever felt disappointed or proud (Warner, Appleby, Whitton, & Faragher, 1997). Targets in the divorce study completed a parallel questionnaire assessing their attitudes and beliefs about their parents' divorce (Laumann-Billings & Emery, 2000), as well as rating their emotions after talking about parental divorce.

The videotapes were then shown to "perceiver" subjects. In the new-motherhood study, perceivers fell into three categories: women who were new first-time mothers of infants, women who were pregnant with their first child, and women who had never been pregnant or raised a child. The three groups of perceivers were roughly similar in terms of age and educational level. In the alcoholism study, perceivers were either self-admitted alcoholics drawn from the same community AA groups as the targets or college students who self-reported that they were not alcoholics. In the divorce study, all participants served as both perceivers and targets, with each student in a pair getting a chance both to talk about their experiences *and* to try to understand their partner's experience.

As the perceivers watched the targets' videos, the experimenter stopped the tapes at the same places that the targets had reported having thoughts and feelings, at which point the perceivers were asked to guess the content of the targets' thoughts and feelings. Perceivers in the new-motherhood and divorce studies were also asked to guess how targets would respond to questionnaires about their respective experiences (as well as to guess how the target felt emotionally about the discussion in the divorce study). Participants in the new-motherhood study, who, as in the alcoholism study, never actually interacted in person with the targets, were asked to write a letter to the target whose video they watched, responding to her experiences, and these letters were later coded by the researchers. Perceivers wrote letters to targets in the alcoholism study, but these letters were only read by the targets.

The three studies also included a third phase during which targets evaluated the perceivers' empathy. Targets coded how accurately the perceivers inferred their thoughts and feelings, read the letters in the new-motherhood and alcoholism studies, and then answered questions about how well they thought the perceivers understood them.

Despite the similarities, the three studies differed in several notable ways. As previously mentioned, targets and perceivers played uniquely different roles and never met face-to-face in the new-motherhood and alcoholism studies, whereas in the divorce study participants had a chance to converse in the first part of the study and both did exactly the same tasks in parallel, rather than one being designated “target” and the other “perceiver.”

Perceivers both with and without similar experiences were drawn from the same general population as targets in the new-motherhood and divorce studies, and thus both kinds of perceivers and targets resembled each other closely on many demographic dimensions. In contrast, in the alcoholism study, “experienced” perceivers came from the same source as targets (local AA chapters) and represented the demographic diversity of such groups, but nonalcoholics (those without the life experience) were drawn from a college student sample and were generally younger than the targets and much less diverse in terms of socioeconomic status and a variety of other variables.

Finally, the three life experiences (new motherhood, alcoholism, and parental divorce) differed markedly in the valence of emotions and attributions generally associated with them. Although all of these life experiences would clearly spark a whole range of emotions, with the particular mix of emotions varying widely across individual cases, the birth of a new child is stereotypically considered a joyous event, whereas the divorce of one’s parents or admitting alcoholism is generally thought to be negative, causing pain and sadness. Furthermore, although the last two events share a negative emotional valence, children are generally not considered responsible for their parents’ divorce (although convincing children of divorce that they are not to blame is often a challenge), whereas alcoholics are often held accountable for their role in abusing alcohol (a judgment that is contested by many).

Despite the differences, there was one clear and strikingly similar result across the three studies: an utter lack of evidence that shared experience improved empathic accuracy. We used three different groups of perceivers in the new-motherhood study, two different kinds of perceivers in the alcoholism study, and two different kinds of dyads in the divorce study (that is, matched pairs in which both participants had divorced parents and unmatched pairs in which only one participant had divorced parents), and collected four different measures of empathic accuracy (detailed below). Across this range of results, perceivers who themselves had experienced the life events that the targets described did no better at getting inside the targets’ heads than perceivers who had not had these experiences.

First among the results, there were no significant differences in “experienced” perceivers’ ability to infer retrospective thoughts and feelings using Ickes’s empathic accuracy paradigm, which utilizes objective coders to rate the accuracy of the perceivers’ inferences about the targets’ thoughts and feelings. Second, using a new variation on Ickes’s method, in all three of the studies we asked the targets themselves to read over the perceivers' inferences and rate them for accuracy. Although targets were generally more generous in their accuracy ratings than the objective coders, their ratings resembled the coders' in that they rated perceivers with and without the experiences as achieving about the same level of accuracy.

As a third measure of accuracy, in the new-motherhood and divorced studies, we examined how well perceivers could guess how the target would respond to scales designed to measure adjustment to these serious life events. New mothers were not significantly more accurate at guessing how other new mothers would respond to the Maternal Attitudes Questionnaire (Warner et al., 1997) when compared to pregnant women or nonmothers (although the trend was at least in the predicted direction for this measure). College students whose parents were divorced were no better than college students whose parents had stayed together at guessing how other students with divorced parents would respond to a questionnaire assessing reactions to divorce (Laumann-Billings & Emery, 2000). As a fourth and final measure of accuracy, in the divorce study we asked perceivers to guess how the targets felt after the discussion of how divorce affects children growing up, using a series of emotion adjectives. Once again, "experienced" perceivers—that is, college students who been through divorce—were no better than inexperienced perceivers at guessing the emotional impact of these discussions on other students who had been through divorce.

Of course, finding no difference between groups constitutes null results, raising the specter that the findings are due to a lack of power. Although distinguishing whether one is dealing with a true case of no difference or merely insufficient power to demonstrate a difference is challenging, we think there are a number of factors pointing toward the former. Consider first that, of all the measures of empathic accuracy (eight different scores collected using four different methods across three studies), the only one for which the means were even in the right direction was perceivers' guesses about how targets would answer the maternal attitudes questionnaire. The scores from Ickes's traditional empathic accuracy paradigm in the alcoholism study were quite strongly trending in the direction that gave nonexperienced (i.e., nonalcoholic) perceivers the edge. If anything, more power in this study would have resulted in a statistically significant advantage for perceivers who had no personal experience being alcoholics!

Even more important, although sample sizes in two of the studies (20 targets and 60 perceivers in the new-motherhood study and 58 pairs of students in the divorce study) were not huge, they were large enough to show some differences between experienced and nonexperienced perceivers on the other measures of empathic response that we collected. The results for empathic concern (i.e., feelings of tenderness, caring, and sympathy toward the other person; see Batson, Early, & Salvarani, 1997), as contrasted with accurately taking someone's perspective, revealed very interesting patterns. In the new-motherhood study, the empathic concern results produced a nice (and statistically significant) stairstep pattern: the more similar the life experience, the more empathic concern perceivers expressed for new-mother targets. New-mother perceivers reported greater empathic concern for targets than pregnant perceivers, who in turn reported greater empathic concern than never-pregnant perceivers.

However, this pattern was reversed for empathic accuracy results in the divorce study. Here, it was participants who did not have the experience (those whose parents were not divorced) who reported feeling the greatest empathic concern for targets whose parents are divorced. The group that expressed the least empathic concern for their discussion partners were the participants whose own parents were divorced but who talked with another participant whose parents were not divorced. (Of course, it could be pointed out that perhaps these participants showed less empathic concern because their partners did not need any—i.e., why work up any concern for someone whose parents stayed together?) The pairs of participants who both had divorced parents showed levels of empathic concern that were in between, but not significantly different from, either of the two extremes found in the unmatched pairs (which were significantly different from each other).

Finally, although experienced perceivers showed no differences on accuracy measures, they were perceived by the targets as differing from inexperienced perceivers. One critical piece of the puzzle that we were interested in examining in the new-motherhood and divorce studies was whether the targets of empathy felt better understood when they interacted with someone who had had the same experience as them.[1] The answer appears to be yes, as long as the targets *knew* that the perceiver had had the experience. In the divorce study, this was virtually always the case. Even though participants were not directly instructed to share their own status regarding parental divorce, in all but two of the dyads this information was revealed in the course of the conversation about divorce.[2]

In the new-motherhood study, perceived understanding was presumably based largely on what perceivers said in the letters they wrote to the targets, given that targets and perceivers never interacted face-to-face. At first glance, the pattern of results for how much targets felt understood appeared to mirror the stairstep results found for empathic concern: the more relevant experience that perceivers had, the more targets reported feeling understood (i.e., targets felt most understood by new-mother perceivers and least understood by women who had never been pregnant or raised a child). However, closer examination revealed that this result was moderated by whether new-mother perceivers also revealed that they were new mothers in their letters. When they did, they were seen as the most understanding of perceivers, but when they didn't reveal that they, too, were mothers, they were rated among the least understanding. For the other two groups of perceivers (pregnant and never-pregnant), revealing one's status had little effect.

Thus, without revealing their special status, new-mother perceivers were not perceived as any more understanding than the other kinds of perceivers. Furthermore, in coding the letters on various dimensions, we found few differences in content among the three groups of perceivers. Similar quantities of advice, compliments, and encouragement were provided by all letter writers.[3] Thus, although all perceivers were largely telling targets the same things in their letters, the targets appeared to interpret the content of the letters differently, depending upon whether the writer reported having had the relevant experience or not.

A related result was also seen in the divorce study, further supporting the idea that a perceiver's response can be viewed differently by the target, depending on what the target thinks the perceiver's experience is. Among target participants who had divorced parents, their partner's reported level of empathic concern differentially predicted how understood the target felt, depending on whether the partner had divorced parents or not. When both participants had divorced parents, greater levels of reported empathic concern on the part of perceivers was associated with the target feeling more understood. However, when divorced target participants interacted with partners whose parents were *not* divorced, these targets reported feeling *less* understood when their partners reported feeling greater empathic concern. The results suggest interacting with someone who feels sorry for you is only comforting when (you think) that person is in the same boat as you. Otherwise, the other person's sympathy can actually hurt the interaction.

Thus, it would appear that shared experience has a greater impact on how the *target* of empathy views the interaction than it does on the behavior of the person trying to be empathic (the perceiver). Experience doesn't make perceivers any more empathically accurate, and it doesn't consistently boost their empathic concern for others who have had the same experience ("experienced" perceivers expressed more empathic concern in the new-motherhood study but not in the divorce study or alcoholism study). Finally, as best we can tell, the effect that experience has on the levels of empathic concern that perceivers report appears to depend on the type of experience that is shared.[4]

The only empathic outcome that was consistently improved by having had a similar experience was the degree to which the target of empathy felt understood. However, our hunch is that a good part of this perceived understanding was driven not by actual differences in the behavior of "experienced" and "inexperienced" perceivers but instead by the targets seeing the perceivers' behavior through different eyes, depending on whether they believed the perceiver had had the experience or not. In other words, the same behavior on the part of the perceivers may be interpreted differently by targets, depending on whether it's coming from someone who has "been there too" or someone who "has no clue." Specifically, partners in the divorce study who expressed greater empathic concern were seen as more understanding only if it was known that they had shared the target's experience. Similarly, in the new-motherhood study, we know that the letters written by the perceivers were the same in terms of length, positivity, and the amount of advice they contained, regardless of which kind of perceiver wrote them. However, the writers of these letters were rated as more understanding when they mentioned in their letter that they had a similar experience: output that was roughly equivalent in objective terms appears to communicate greater understanding if it comes from someone who has had the experience.

It is notable that in the parental divorce study, where everyone was both target and perceiver, almost everyone provided information about their status regarding this experience, often because they were asked by the other participant. In the new-motherhood study, where revealing this information was entirely up to the perceivers, only about half the perceivers chose to do so. We can tentatively conclude that getting information about shared experience status is more important to targets than revealing such information is to perceivers.

The results of these studies suggest that knowledge of the other person's similar experience has at least as much (if not greater) influence on the *target* of empathy as the actual experience has on the empathic response of the perceiver. Whereas past research has mostly focused on the *perceiver's* empathic accuracy and empathic concern in social interactions, the present research on similar experience reminds us that the alignment between the two minds that characterizes empathy goes both ways.

What are the practical applications of these findings? First, and obviously, there appear to be sizable benefits in letting other people know when you have had similar life experiences. This should come as no surprise to social psychologists—who have known for years about the importance of similarity in creating positive social regard (Byrne, 1961)—or presidential candidates—who spin their upbringing a dozen ways in order to always be able to claim that they have come from the same background as their constituents. Given that the benefit seems to come largely from simply being perceived as having had a similar experience, unaccompanied by other apparent changes in behavior, if ethical concerns are set aside, there may well be advantages to merely claiming to have had a particular life experience.

Future research may identify moderators that can change the general conclusions that we have drawn from these three studies. One possible moderator is valence of experience. We found that shared experience increased the perceiver's empathic concern when the experience was new motherhood (generally considered a positive experience) but not when the experience was parental divorce (generally considered a negative experience). Does valence of experience consistently moderate empathic concern results in this way? Another variable that may affect results is whether the experience is something people are seen as having brought on themselves (e.g., alcoholism) or something that happens to them (e.g., parental divorce).

In addition, there may be other circumstances not covered by these three studies which *do* produce consistent effects of experience on accuracy and other measures of empathy. Our participants had very constrained interactions with each other (limited to about 5 minutes in the divorce study and limited to non-face-to-face communication in the other two studies). Less structured and more extensive interactions may produce different results. For example, with more in-depth interactions, the wisdom that comes with experience may start to pay off in terms of empathic accuracy. Another possibility is that, with more contact, *other* measures of empathic response will grow to look more like accuracy does under limited contact—that is, these measures will also be leveled. For example, empathy targets may come to revise their initial assumption that "experienced" perceivers understand them better; perceivers may start to feel levels of empathic concern that are based on the details of the target's story (and how the target tells that story) rather than simply on whether the target has had an experience or not.

Perhaps most importantly, future research needs to consider both the actual and the perceived range of variation in the experience. Shared experience may have more powerful effects when the experience is thought to be one that is the same across people, whereas these effects may be dampened for experiences that are thought to vary greatly from person to person (e.g., one's first plane ride). Additionally, it may matter whether people can clearly agree whether they have both had a particular life experience or not. For some experiences, such as being the victim of abuse, two individuals may not agree that they have had a common experience (e.g., what one person may think is spousal abuse may be considered a normal part of marriage to another person). In the case of new motherhood and parental divorce, agreement is likely to be higher, but even with experiences such as these, where the distinction between those who "have" and those who "have not" is clear, the range of participants' supposedly "similar" experiences was striking. New-motherhood experiences in our study ranged from the newly single mother who discovered she was unexpectedly pregnant right after she and her husband had decided to divorce to the woman whose new-motherhood experience came right out of fairy tale, finding that a desired and planned pregnancy brought her even closer to the husband she already adored. Divorce tales included a touchingly dazed college freshman whose parents (married for over 20 years) told him during the fall term of his first year of college that they were divorcing, but also a student whose parents divorced before she was born and who acquired a much beloved stepfather in her preschool years. Both experiences were different from the several targets who experienced a lot of parental fighting before seeing their parents divorce in their early teens.

With such variations, we might very well be able to find someone in the group of students with still-married parents who had more in common with each of these two examples than the two had in common with one another. In other words, variation within a particular experience category—at least for some life experiences—may be as great as or greater than the variation across categories of experience. I would argue further that the level of within-category variation may have a greater effect on the empathic perceiver than the target of empathy. Why? Because only the perceiver has to face the "other minds" problem; the target of empathy does not. When people find themselves trying to know what is going through another person's mind, they can't directly comprehend his or her subjective experience. They have to try to construct a mind from a set of incomplete and sometimes fuzzy clues. One possible resource for filling in the blanks is "self" experience—one's own experience (Dawes, 1990; Hodges, Johnsen, & Scott, 2002). Thus, we might expect that variations in selves will carry over into variations in comprehension of the other.

However, when someone is the target of empathy, he or she can just skip the problem of trying to comprehend the ever elusive "other mind" (although people who are particularly attributionally complex, highly neurotic, or fascinated by social psychology are probably unable to do this!). A target can simply look within and ask, "Do I feel understood?" In answering this question, people may consult shared beliefs or expectations, such as the (apparently!) ingrained idea in our culture that shared experience makes others more understanding. Another possibility, one that I think merits future research, is that targets may give the benefit of the doubt to someone else who has had a similar experience because that experience makes that person a member of the target's ingroup.

In sum, I started out with what I thought was a simple, easy-to-answer question: Does experience affect empathy? The attempts by my colleagues and me to answer this question have yielded results that are surprising, inconsistent, and incomplete. Our only solace is that researchers like us who have explored this tricky domain will *surely* understand our frustration and fascination!

NOTES

1. There were no significant differences in perceived understanding (or empathic concern) in the alcoholism study, quite conceivably due to small sample sizes. Interestingly (and consistent with the results that follow in this section), the alcoholic targets rated the alcoholic perceivers as (nonsignificantly) most understanding overall.
2. The revelation tended to come sooner rather than later as well. In 25% of the dyads, both participants had shared their parental divorce status in the first 12 seconds of the interaction, and 30 seconds into the interaction both participants had revealed their status in 50% of the dyads. There were only two dyads (both pairs of males) in which both participants did not reveal their status. In one, the interaction took the form of one participant with divorced parents revealing this information fairly early on and the other participant (who never stated that his parents weren't divorced) asking questions about the first participant's experience. In the only other dyad, neither participant revealed their parental divorce status early on. Even after one finally did, the conversation continued to be among the most—if not *the* most—awkward and stilted in the entire study, in which conversations were generally smooth and animated.
3. Where differences occurred, they centered around numbers of comments about how similar or different the perceiver was to the target, which often included the revelation of one's own motherhood status (e.g., "I believe when I have my first child I will be completely frazzled," or "[My baby] is almost sleeping through the night"). Interestingly, new-mother perceivers made notably more comments about both how similar to, *and* different from, the targets they were.
4. Dan Batson and colleagues (1996) have also examined the relationship between experience and empathic concern and found an interesting sex difference: women, but not men, reported greater empathic concern with shared experience. Similarly, in the new-motherhood study (involving only female participants), we also found experience producing greater empathic concern. In the divorce study, neither men nor women showed greater empathic concern with experience.

REFERENCES

Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective taking: Imagining how another feels versus imagining how you would feel. *Personality and Social Psychology Bulletin, 23,* 751–758.

Batson, D. C., Sympson, S. C., Hindman, J. L., Decruz, P., Todd, R. M., Weeks, J. L., et al. (1996). "I've been there, too": Effect on empathy of prior experience with a need. *Personality and Social Psychology Bulletin, 22,* 471–482.

Byrne, D. (1961). Interpersonal attraction and attitude similarity. *Journal of Abnormal and Social Psychology, 62,* 711–715.

Dawes, R. M. (1990). The potential nonfalsity of the false consensus effect. In R. M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 171–199). Chicago: University of Chicago Press.

Hallinan, B. (2000). *Experience, empathic accuracy, and alcoholism*. Unpublished honors thesis, University of Oregon, Eugene.

Hodges, S. D. (2004, January). You just can't understand: Shared experience and parental divorce. In J. A. Simpson (Chair), *The perils of accuracy and inaccuracy in relationships and social interactions*. Symposium conducted at the annual meeting of the Society of Personality and Social Psychology, Austin, TX.

Hodges, S. D., Johnsen, A. T., & Scott, N. S. (2002). You're like me, no matter what you say: Self projection in self-other comparisons. *Psychologica Belgica, 42,* 101–112.

Hodges, S. D., Klein, K. J. K. K., Veach, D., & Villanueva, B. R. (2004). *Giving birth to empathy: The effects of similar experience on empathic accuracy, empathic concern, and perceived empathy.* Unpublished manuscript, University of Oregon, Eugene.

Ickes, W. (2001). Measuring empathic accuracy. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 211–241). Mahwah, NJ: Erlbaum.

Laumann-Billings, L., & Emery, R. E. (2000). Distress among young adults from divorced families. *Journal of Family Psychology, 14,* 671–687.

Warner, R., Appleby, L., Whitton, A., & Faragher, B. (1997). Attitudes toward motherhood in postnatal depression: Development of the Maternal Attitudes Questionnaire. *Journal of Psychosomatic Research, 43,* 351–358.

20

Empathic Accuracy and Inaccuracy in Close Relationships

WILLIAM ICKES
JEFFRY A. SIMPSON
MINDA ORIÑA

SHE: I think we should talk.
HE: What do you want to talk about?
SHE: How about what's-her-name, your new "friend"?
HE: Let's not go there.
SHE: Yes, let's go there.

As desirable as the truth may be, the harsh reality is that sometimes the truth hurts. Indeed, in some situations, people may not be able to handle the truth and might even be motivated to avoid it altogether. In other situations, however, people may realize that knowing the truth will hurt but still attempt to seek it out, even if the resulting knowledge might prove painful. Accordingly, whenever we seek to "read" other people's minds, we run the risk of discovering some unpleasant truths that may have the potential to alter our relationships.

Our goal in this chapter is to explore how the potential for discovering such unpleasant truths affects our empathic accuracy in long-term established relationships. To impose some theoretical rigor on our pursuit of this goal, we consider the motivational implications of *empathic accuracy*—the degree to which individuals accurately infer their relationship partners' thoughts and feelings—in terms of a general rule, its major exception, and the logical complement of that exception.

The general rule suggests that relationship partners who accurately infer each other's benign, nonthreatening thoughts and feelings should be more successful at maintaining satisfying, stable relationships than less accurate partners are. The major exception to this rule suggests that the opposite is true when the partners' thoughts and feelings are relationship-threatening (i.e., greater accuracy in this case can often hurt the partners' relationship). Finally, the logical complement of this exception to the rule is that *motivated inaccuracy* can help relationships in those cases when knowledge of the partner's relationship-threatening thoughts and feelings would otherwise destabilize them.

The general rule, its major exception, and the logical complement of that exception have been conceptually integrated within the framework of our empathic accuracy model (Ickes & Simpson, 1997, 2001). Thus, we begin by describing the empathic accuracy model, which describes how individuals attempt to "manage" empathic accuracy within their close relationships. We then review a program of research that has provided provisional support for the major tenets of the model (for more detailed overviews, see Ickes, 2003, and Ickes & Simpson, 2001).

THE EMPATHIC ACCURACY MODEL

Conventional wisdom suggests that understanding a relationship partner's thoughts and feelings should typically be beneficial for relationships. Empirical research has revealed, however, that this widely held belief is too simplistic (see Sillars & Scott, 1983). Although greater empathic accuracy is often associated with greater relationship satisfaction and stability in situations that pose little or no threat to relationships (e.g., Kahn, 1970; Noller, 1980), it is associated with *less* satisfaction and *less* stability in relationship-threatening situations (e.g., Sillars, Pike, Jones, & Murphy, 1984; Simpson, Ickes, & Blackstone, 1995).

At first blush, these findings seem counterintuitive when one considers that threats to relationships might be more easily defused or better resolved if partners are able to understand each other's thoughts and feelings more accurately. To resolve this apparent paradox, we developed a theoretical model that specifies how relationship partners might "manage" their levels of empathic accuracy in relationship-threatening versus nonthreatening situations (Ickes & Simpson, 1997, 2001). More specifically, the model identifies conditions under which (1) empathic accuracy can help relationships (the general rule); (2) empathic accuracy can hurt relationships (the major exception to the rule); and (3) empathic *in*accuracy can help relationships (the logical complement of the exception to the rule).

The model begins with the assumption that the potential upper and lower limits of empathic accuracy within a given interaction are determined by two factors: (1) the partners' respective levels of "readability" (i.e., the degree to which each partner displays cues that reflect his or her true internal states), and (2) the partners' respective levels of empathic ability (i.e., the degree to which each partner can accurately decipher the other's valid behavioral cues). Within these boundaries, however, empathic accuracy should be "managed" very differently, depending on the nature of the situated interactions in which the partners find themselves. The types of interactions that are most central to our model are represented in Figure 20.1.

In Figure 20.1, the individual (rather than the dyad) is treated as the primary level of analysis. According to our model, when relationship partners enter a situation, each individual first determines whether the current situation is likely to present a *danger zone* for the relationship. The danger zones for relationships are insights that would threaten the relationship if the individual perceivers accurately inferred their partner's actual relationship-threatening thoughts and feelings in that particular situation.

At the first branching point of the model, the perceiver decides whether or not a danger zone issue might be present (or is likely to emerge) in the current situation. Let us first consider the portion of the model that pertains to situations in which individuals perceive that the current situation is nonthreatening (i.e., that no danger zone issues are likely to emerge that could force the partners to deal with each other's relationship-threatening thoughts and feelings).

EMPATHIC ACCURACY IN NONTHREATENING CONTEXTS

When perceivers believe that they will discuss issues that are not likely to be relationship-threatening (see the right side of Figure 20.1), the model predicts that they should be motivated to infer their partner's thoughts and feelings accurately, that personal and relational distress should be low, and that their relationship should remain stable. The reasoning behind these predictions is relatively straightforward. To the extent that (1) mutual understanding typically facilitates the coordination of joint actions so that both personal and relational goals can be better achieved,

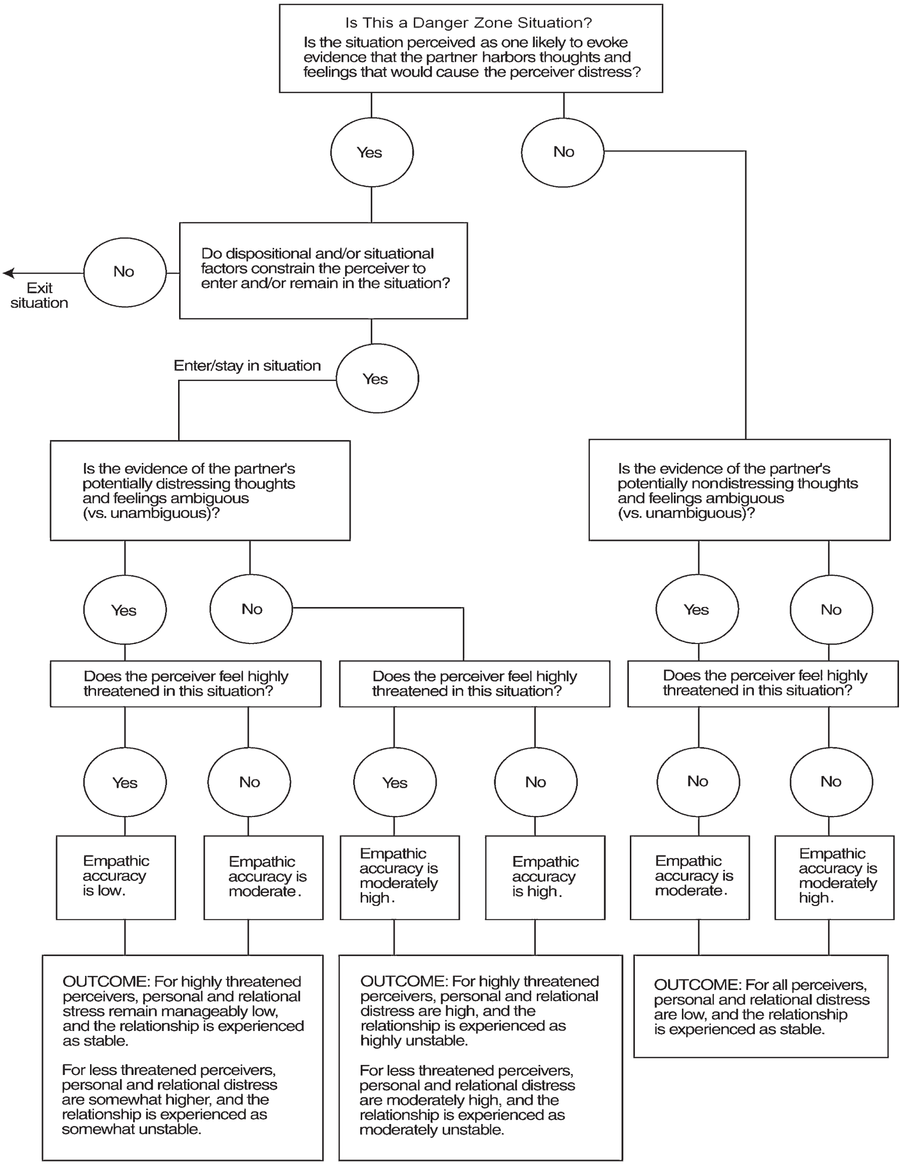
Is This a Danger Zone Situation?
Is the situation perceived as one likely to evoke evidence that the partner harbors thoughts and feelings that would cause the perceiver distress?
Yes
No
Do dispositional and/or situational factors constrain the perceiver to enter and/or remain in the situation?
No
Exit situation
Enter/stay in situation
Yes
Is the evidence of the partner's potentially distressing thoughts and feelings ambiguous (vs. unambiguous)?
Is the evidence of the partner's potentially nondistressing thoughts and feelings ambiguous (vs. unambiguous)?
Yes
No
Yes
No
Does the perceiver feel highly threatened in this situation?
Does the perceiver feel highly threatened in this situation?
Does the perceiver feel highly threatened in this situation?
Yes
No
Yes
No
No
No
Empathic accuracy is low.
Empathic accuracy is moderate.
Empathic accuracy is moderately high.
Empathic accuracy is high.
Empathic accuracy is moderate.
Empathic accuracy is moderately high.
OUTCOME: For highly threatened perceivers, personal and relational stress remain manageably low, and the relationship is experienced as stable.
For less threatened perceivers, personal and relational distress are somewhat higher, and the relationship is experienced as somewhat unstable.
OUTCOME: For highly threatened perceivers, personal and relational distress are high, and the relationship is experienced as highly unstable.
For less threatened perceivers, personal and relational distress are moderately high, and the relationship is experienced as moderately unstable.
OUTCOME: For all perceivers, personal and relational distress are low, and the relationship is experienced as stable.

FIGURE 20.1. The empathic accuracy model. Adapted from previous versions of the model published in Ickes and Simpson (1997, 2001).

and (2) the behaviors needed to achieve accurate understanding tend to be reinforced over time, most perceivers should be disposed to achieve moderately high levels of empathic accuracy in non-relationship-threatening situations.

As a result, in situations where no danger zones are perceived as imminent (e.g., during everyday conversations about benign or nonthreatening issues), most perceivers should display a habit-based "accuracy" orientation that should enable them to clarify minor misunderstandings, keep routine conflicts from escalating into major ones, and gain a deeper understanding of their partner's stance on these issues. These tendencies, in turn, should sustain or even enhance relationship satisfaction and stability. Thus, as long as situations do not threaten perceivers with the emergence of danger zone issues that could reveal a partner's relationship-threatening thoughts and feelings, most perceivers should achieve at least moderate levels of empathic accuracy, with the ambiguity versus "diagnosticity" of the partner's behavior accounting for much of the variability in this case (see the middle-right portion of Figure 20.1).

Another factor influencing this variability is, of course, the perceiver's motivation to be accurate, which may often be attenuated by the routine, taken-for-granted nature of these nonthreatening interactions (Thomas, Fletcher, & Lange, 1997). For this reason, the perceiver's empathic accuracy in nonthreatening situations should be moderate rather than high (see the lower-right side of Figure 20.1). In general, however, higher levels of empathic accuracy should result in higher levels of relationship satisfaction and stability in nonthreatening situations, consistent with the rule that *greater empathic accuracy should generally help relationships* (i.e., in situations that do not threaten the relationship).

EMPATHIC ACCURACY IN RELATIONSHIP-THREATENING CONTEXTS

Inevitably, individuals encounter situations in which danger zone topics or issues might emerge and threaten to destabilize their relationships (see the left-hand side of Figure 20.1). When such relationship-threatening situations arise, our model predicts that most perceivers' initial impulse should be to avoid or escape from these situations, if possible. In other words, avoiding or escaping from danger zone situations will typically be the first line of defense that perceivers use to "manage" their empathic accuracy, as it allows them to avoid having to confront their partners' potentially relationship-damaging thoughts and feelings.

The use of this strategy presumes, of course, that perceivers can recognize—and perhaps even anticipate—potential danger zone issues in their relationships. With time, perceivers in most, if not all, relationships will gradually learn to identify and avoid these issues in order to protect their own self-esteem, their partner's self-esteem, and/or their cherished views of the relationship. By doing so, perceivers can avoid dealing with these danger zone topics altogether, acting as though it is better (and easier) to avoid confronting one's worst fears than it is to have one's worst fears confirmed and then being forced to deal with them.

Needless to say, avoiding or escaping danger zone issues is not always possible (see the left and middle portions of Figure 20.1). When perceivers must remain in a relationship-threatening situation, the model posits that their second line of defense will be *motivated inaccuracy*—a conscious or unconscious failure to accurately infer the content of their partner's potentially harmful thoughts and feelings. The ultimate success of this strategy should hinge on the degree to which the cues to the partner's potentially threatening thoughts and feelings are ambiguous versus unambiguous.

If the cues to the partner's potentially threatening thoughts and feelings are ambiguous (see the middle-left side of Figure 20.1), perceivers should be able to use motivated *in*accuracy as a defense. By using *subception* (Carl Rogers's term for tuning out, distorting, or reframing potentially threatening information) and other psychological defense mechanisms (e.g., *denial*, *repression, rationalization*) that shelter individuals from recognizing the most threatening implications of their partners' potentially destructive thoughts and feelings, perceivers should end up displaying low levels of empathic accuracy.[1] These defenses should at least temporarily benefit both perceivers and their relationships by minimizing personal and relational distress and by keeping these relationships stable in the face of potentially destabilizing forces. Accordingly, the left-hand portion of the model illustrates the logical complement of the major exception to the general rule—that is, that *motivated inaccuracy can help to sustain relationships in the face of threat.*

What happens when individuals (perceivers) remain in relationship-threatening situations but cannot use motivated inaccuracy as a secondary strategy for curtailing the threat? The middle section of Figure 20.1 highlights this case. When the cues to the relationship-threatening content of the partner's thoughts and feelings are clear and unambiguous (e.g., the partner openly admits that he or she is having an affair with someone else), the stark clarity of this information should force most perceivers to achieve at least moderately high levels of empathic accuracy, accompanied by sharp declines in both relationship satisfaction and stability. This is a clear instance in which greater empathic accuracy should harm relationships. However, because the unwilling perceiver is forced to be accurate in response to the sheer clarity of the information, it is not a situation in which the perceiver's *motivated* accuracy harms relationships.

The special case we call motivated accuracy occurs only when the individual perceiver possesses a strong personal need or disposition to confront the truth about a partner's relationship-relevant thoughts and feelings. This case is not depicted in the simplified model presented in Figure 20.1. However, according to the latest elaborations of our model, as described in written form by Ickes and Simpson (2001) and by Ickes (2003), need- and disposition-based accuracy motives will at times override the inclination to either avoid danger zone topics and issues entirely or to use motivated inaccuracy to diffuse potential relationship threats. Instances of this type constitute a special case of the major exception to the general rule—that is, that *motivated accuracy can also hurt relationships when the partners' thoughts and feelings are relationship-threatening,* just as unmotivated (i.e., situationally constrained) accuracy can.

Finally, there is another special case that is implied by, but not formally represented in, our Figure 20.1 model. It occurs when the perceiver accurately "reads" the partner's relationship-threatening thoughts and feelings, experiences them as undermining the relationship to the extent that it is no longer viable, but eventually finds some degree of solace by feeling "sadder but wiser" (and thereby ultimately better off) for the insights that have been gained. Because the solace gained in this case is often far removed in time from the highly distressing events that were immediately triggered by the perceiver's empathic accuracy, this special case lies outside the more immediate process model that Figure 20.1 aspires to represent. It is important to acknowledge, however, that even the most painful and disruptive of our empathic insights have the potential to transform into a healing and adaptive wisdom later on.

PRELIMINARY TESTS OF THE EMPATHIC ACCURACY MODEL

We have tested portions of the empathic accuracy model in a series of social interaction studies involving romantic couples (both dating and married couples). In all of these studies, we have used Ickes, Bissonnette, Garcia, and Stinson's (1990) procedures for measuring the partners' "online" empathic accuracy (for more details about these procedures, see Ickes, 2001; Ickes, Stinson, et al., 1990).

Evidence for the Model's Basic Interaction Prediction: Testing the General Rule and Its Major Exception

According to the empathic accuracy model, when a partner's thoughts and feelings are relatively benign and nonthreatening, greater empathic accuracy by the perceiver should be associated with an increase in the perceiver's feeling of closeness to the partner (i.e., relative to the perceiver's feeling before empathic accuracy increased). This prediction reflects the general rule that a better understanding of one's partner's benign and nonthreatening thoughts and feelings should be associated with positive relationship outcomes. On the other hand, when a partner's thoughts and feelings are relationship-threatening, our model predicts that greater empathic accuracy on the part of the perceiver should be associated with a decline in the perceiver's feeling of closeness. This latter prediction reflects the major exception to the rule—that when a partner's thoughts and feeling are relationship-threatening, greater empathic accuracy should be associated with negative relationship outcomes.

The interaction effect implied by these contrasting predictions was recently tested in a study of 96 married couples (Simpson, Oriña, & Ickes, 2003). We recruited the couples (mean length of marriage = 5.8 years) to participate in a videotaped conflict resolution task. After being separated, the partners independently completed a set of questionnaires that asked about their backgrounds and their marriage. They also reported their current (prediscussion) feeling of closeness to their spouse. The partners were then reunited and asked to identify and "try to resolve" an unresolved problem in their marriage, one that concerned a relationship-threatening issue. Each couple was given 7–10 minutes to resolve their specified problem as best they could.

Immediately after the discussion, each spouse was escorted to a private room where he or she watched the videotaped interaction and listed all of the thoughts and feelings that he or she had at various time points during the interaction. During a second viewing of the videotape, each spouse then completed the empathic inference task, during which he or she tried to infer the content of each of his or her *partner's* thoughts and feelings at each of the time points listed by the partner. Finally, each spouse completed a set of postinteraction measures, which included a postinteraction assessment of his or her current feeling of closeness to the partner.

The significant interaction predicted by the empathic accuracy model is precisely what we found. The more accurately the perceivers inferred their partner's thoughts and feelings when their partners actually harbored relationship-threatening thoughts and feelings, the *less close* the perceivers reported feeling to their partners after the conflict discussion, relative to what they had reported feeling before the discussion. Conversely, the more accurately the perceivers inferred their partner's actual benign and nonthreatening thoughts and feelings, the *closer* the perceivers reported feeling to their partners following the conflict discussion than they had reported feeling before it. This interaction effect remained significant even when the perceivers' self-reported and observer-rated levels of relationship-threatening thoughts and feelings were statistically controlled.

Testing the Major Exception's Logical Complement: When Motivated Inaccuracy Helps Relationships

One limitation of the Simpson, Oriña, and Ickes (2003) study is that it does not provide a direct test of the logical complement of the major exception to the rule—that there are circumstances in which motivated *in*accuracy can help relationships by protecting them from the destabilizing influences of a potential threat. Fortunately, a 1995 study by Simpson, Ickes, and Blackstone provided such a test. In that study, we asked long-term heterosexual dating partners (mean length of relationship = 16.5 months) to take turns rating and discussing with their partner the physical and sexual attractiveness of opposite-sex persons who were described as "potential dating partners." About half of the couples were randomly assigned to view and rate slides of highly attractive people, whereas the remaining couples viewed and rated slides depicting less attractive people. After the male (or the female) partner rated aloud the attractiveness and sexual appeal of each opposite-sex stimulus person on a 10-point rating scale, both partners then discussed what they liked or disliked about each person for approximately 30 seconds. Each couple's interaction was covertly videotaped.

Immediately following the slide rating task, both partners were told about the taping and were asked to give their written consent to allow us to view and code their videotape. They were also asked to participate in the next phase of the study, which involved having each partner complete the empathic accuracy rating task. If they agreed, the dating partners were then separated and escorted to different rooms, where each partner privately viewed a copy of the videotape of their interaction throughout the entire course of the rating-and-discussion task. Each partner made a written record of each of his or her thoughts and feelings and the "tape stops" (i.e., the specific times during the interaction) at which each thought or feeling occurred. Each partner then viewed the videotape again, this time attempting to infer as accurately as possible the specific content of each of their *partner's* thoughts and feelings at their respective "tape stops." The dating partners were then released from the study, no doubt thinking it was over. Four months later, however, both partners were contacted by telephone to determine whether or not they were still dating each other.

The main purpose of this study was to determine whether certain dating partners would exhibit motivated *in*accuracy and thereby minimize the relational instability and dissatisfaction that could have resulted from accurately inferring their partners' generally favorable (and relationship-threatening) thoughts about highly attractive opposite-sex people. As predicted, we found that the *lowest* empathic accuracy during the rating-and-discussion task was displayed by dating partners who were closer (i.e., more interdependent), who were less certain about the long-term stability of their relationship, and who rated attractive (vs. less attractive) "potential dating partners" in each other's presence. (For the details of these analyses, see Simpson et al., 1995.)

We also found that the relation between these three variables and empathic accuracy was mediated by the perceiver's level of self-reported threat during the study. That is, the dating partners who were the most interdependent, the least certain about the stability of their relationship, and who rated highly attractive others reported feeling most threatened, and these high levels of threat in turn predicted their objectively poor (i.e., near-chance) levels of empathic accuracy. Even more impressive, the partners in this category were all still dating at 4-month follow-up, whereas the remaining couples in the study had a significantly higher baseline breakup rate of 28%.

In sum, the pattern of perceptions and behavior displayed by these close-but-uncertain partners who rated and discussed attractive others provides solid evidence for the logical complement of the major exception to the rule—that is, that *sometimes motivated inaccuracy actually helps relationships*. Apparently, relationship partners can use motivated inaccuracy in relationship-threatening situations to protect themselves and their relationships from the pain and injury that would otherwise result from a more accurate understanding of what their partner is thinking or feeling. (For a detailed discussion of the entire pattern of evidence that supports our motivated inaccuracy interpretation, see Simpson et al., 1995.)

Testing the Special-Case Exception to the General Rule: When Motivated Accuracy Hurts Relationships

The special-case exception to the general rule suggests that strong accuracy motives can override perceivers' inclinations to either avoid inferring or to misinfer a partner's relationship-threatening thoughts and feelings, thereby hurting the relationship. A study by Simpson, Ickes, and Grich (1999) has shown how such dispositionally based *accuracy motives* can result in motivated accuracy as opposed to motivated inaccuracy.

People who are anxiously attached worry a great deal about their partners leaving or abandoning them (Hazan & Shaver, 1987). For this reason, highly anxious individuals should be hypervigilant to any signs that their partners might be harboring relationship-threatening or relationship-destabilizing thoughts and feelings (see Cassidy & Berlin, 1994; Mikulincer & Florian, 1998). Reanalyzing the data from the 1995 study using the dating partners' scores on the adult attachment dimensions as predictor variables, Simpson and colleagues (1999) found that highly anxious dating partners—particularly highly anxious women—did not display the same motivated inaccuracy during the relationship-threatening slide rating task that was observed in dating partners who had other attachment orientations. On the contrary, the more anxiously attached the women were, the more accurately they inferred their partners' relationship-threatening thoughts and feelings across both of the threat levels (mild and strong) that were experimentally manipulated in this study.

There was a price to be paid for this accuracy, however. For the more anxiously attached the women were, the more they felt distressed, threatened, and jealous, and the more they displayed other signs of acute relationship dissatisfaction and instability. The behavior of these highly anxious women therefore provides some intriguing preliminary evidence for the "special-case exception to the rule"—that is, that *motivated accuracy can sometimes harm close relationships*. (For other findings that are consistent with this view, see Ickes, Dugosh, Simpson, & Wilson, 2003.)

CONCLUSION

Conventional wisdom promotes greater understanding as a sovereign cure for the various ills that can plague close relationships. As we have seen, however, conventional wisdom is, at best, overly simplistic in this case and, at worst, downright misleading. Indeed, in certain situations, the prescription for greater understanding may be a prescription for relationship disaster. The theory and research described in this chapter have taken us beyond conventional wisdom, and the practical implications of this advance are important. For example, the various insights we have gained should help marital therapists and other counselors and clinicians to determine when it would be more adaptive for individuals to seek greater empathic accuracy, and when it would be more adaptive for them to avoid inquiring too closely into their relationship partners' thoughts and feelings. Our findings should also help therapists to identify which cognitive and behavioral interventions may be most appropriate for treating certain types of distressed individuals and couples.

In this chapter, we have used the empathic accuracy model to help specify the situations in which greater empathic accuracy actually does help relationships. More interestingly, however, we have also used the model to specify other situations in which this commonsense connection does *not* hold true. In these latter situations, we have found evidence of two important exceptions to the general rule: (1) cases in which motivated *in*accuracy actually helps relationships, and (2) cases in which strong accuracy motives override perceivers' defenses against processing threatening information, thereby harming their relationships. Although relationship partners often claim that they want to apprehend the "truth" of each other's private thoughts and feelings, there are times when confronting that "truth" can be highly distressing and can incur a very high price. Knowing when to pay this price, and when to avoid paying it, may be one of the most important aspects of "managing" and maintaining close relationships.

NOTE

[1] The processes involved here are complex, controversial, and seemingly paradoxical, as they require the perceiver to first identify a potential threat at a level below conscious awareness and then consciously interpret it in a way that renders it nonthreatening. Our colleagues who model neural networks assure us, however, that such seemingly paradoxical forms of information processing can and do occur, and that they can be modeled successfully using what are commonly termed "backward propagation algorithms."

REFERENCES

Cassidy, J., & Berlin, L. J. (1994). The insecure/ambivalent pattern of attachment: Theory and research. *Child Development, 65*, 971–991.

Hazan, C., & Shaver, P. R. (1987). Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology, 52*, 511–524.

Ickes, W. (2001). Measuring empathic accuracy. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 219–241). Mahwah, NJ: Erlbaum.

Ickes, W. (2003). *Everyday mind reading: Understanding what other people think and feel.* Amherst, NY: Prometheus Books.

Ickes, W., Bissonnette, V., Garcia, S., & Stinson, L. (1990). Implementing and using the dyadic interaction paradigm. In C. Hendrick & M. Clark (Eds.), *Review of personality and social psychology: Vol. 11. Research methods in personality and social psychology* (pp. 16–44). Newbury Park, CA: Sage.

Ickes, W., Dugosh, J. W., Simpson, J. A., & Wilson, C. (2003). Suspicious minds: The motive to acquire relationship-threatening information. *Personal Relationships*, *10*, 131–148.

Ickes, W., & Simpson, J. A. (1997). Managing empathic accuracy in close relationships. In W. Ickes (Ed.), *Empathic accuracy* (pp. 218–250). New York: Guilford Press.

Ickes, W., & Simpson, J. A. (2001). Motivational aspects of empathic accuracy. In G. J. O. Fletcher & M. Clark (Eds.), *The Blackwell handbook in social psychology: Interpersonal processes* (pp. 229–249). Oxford, UK: Blackwell.

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology, 59*, 730–742.

Kahn, M. (1970). Nonverbal communication and marital satisfaction. *Family Process, 9*, 449–456.

Mikulincer, M., & Florian, V. (1998). The relationship between adult attachment styles and emotional and cognitive reactions to stressful events. In J. A. Simpson & W. S. Rholes (Eds.), *Attachment theory and close relationships* (pp. 143–165). New York: Guilford Press.

Noller, P. (1980). Misunderstandings in marital communication: A study of couples' nonverbal communication. *Journal of Personality and Social Psychology, 39*, 1135–1148.

Sillars, A. L., Pike, G. R., Jones, T. S., & Murphy, M. A. (1984). Communication and understanding in marriage. *Human Communication Research, 10*, 317–350.

Sillars, A. L., & Scott, M. D. (1983). Interpersonal perception between intimates: An integrative review. *Human Communication Research, 10*, 153–176.

Simpson, J. A., Ickes, W., & Blackstone, T. (1995). When the head protects the heart: Empathic accuracy in dating relationships. *Journal of Personality and Social Psychology, 69*, 629–641.

Simpson, J. A., Ickes, W., & Grich, J. (1999). When accuracy hurts: Reactions of anxiously-attached dating partners to a relationship-threatening situation. *Journal of Personality and Social Psychology, 76,* 754–769.

Simpson, J. A., Oriña, M. M., & Ickes, W. (2003). When accuracy hurts, and when it helps: A test of the empathic accuracy model in marital interactions. *Journal of Personality and Social Psychology, 85,* 881–893.

Thomas, G., Fletcher, G. J. O., & Lange, C. (1997). On-line empathic accuracy in marital interaction. *Journal of Personality and Social Psychology, 72*, 839–850.

21

Theory of Mind in Schizophrenia

ROBYN LANGDON

A healthy social life depends, among other things, on being sensitive to the subjective lives of other people. Being sensitive to others in this way requires more than the knowledge that other people have minds of their own; we also need to be able to make fairly accurate inferences concerning what another person is likely to be thinking in his or her particular circumstances. A lack of mind-awareness (i.e., mindblindness) will seriously curtail social life, but so too will poor mindreading in individuals who are in principle mind-aware. Indeed, the latter combination may lead to even more serious confusion over the true motives of other people. That mind-awareness and accuracy in mindreading dissociate accords with the findings of research investigating theory of mind in autism versus schizophrenia.

Premack and Woodruff (1978) coined the term "theory of mind" to refer to a capacity to impute causal mental states in order to predict and/or explain behavior. The empirical standard used traditionally to assess whether an individual has a fully developed theory of mind (or a full appreciation of the representational nature of mental life) is a demonstrated understanding that intentional agents can act on the basis of beliefs that misrepresent the true state of affairs. In the classic Sally–

Anne task, for example, a young child is shown two dolls: Sally and Anne. Sally and Anne are depicted playing together. Sally has a marble that she puts in a basket before leaving the room. While she is out, Anne moves the marble to a box. Then she, too, leaves the room. When Sally returns, the child is asked, "Where will Sally look for her marble?" In order to pass this task, the child must represent Sally's false belief that her marble is in the basket.

AUTISM AND THEORY OF MIND

Since the emergence of the theory-of-mind hypothesis of autism over 15 years ago, autistic symptomatology has been considered to be the psychopathological consequence of a lack of appreciation of minds. Autism is an early-onset neurodevelopmental disorder characterized by the "presence of markedly abnormal or impaired development in social interaction and communication and a markedly restricted repertoire of activities and interests" (American Psychiatric Association, 1994, p. 68). Two to five people per 10,000 are affected, and the disorder is four to five times more prevalent in males than in females. The social impairments and the restricted and repetitive patterns of behavior must be evident prior to 3 years of age in order to meet the diagnostic criteria for autism.

Individuals with autism have been consistently reported to perform poorly on theory-of-mind tasks, yet show preserved capacities in other domains (see Baron-Cohen, 1995, for a review). For example, these individuals fail to appreciate the distinction between beliefs and reality, yet understand that photographs can misrepresent current reality (Charman & Baron-Cohen, 1995; Leekam & Perner, 1991). Autistic individuals, while insensitive to other people's beliefs, demonstrate an understanding of other people's basic emotions (e.g., happiness and sadness[1]; Baron-Cohen, 1991; Baron-Cohen, Spitz, & Cross, 1993). Finally, autistic individuals who fail to appreciate *knowing* nevertheless understand *seeing* (Baron-Cohen & Goodhart, 1994; Leslie & Frith, 1988); these individuals demonstrate an understanding that different observers see the same object in different ways, depending on how each observer is spatially located relative to the seen object (Reed, 1994; Reed & Peterson, 1990; Tan & Harris, 1991), even though they fail to appreciate that beliefs are constrained by access to knowledge in a similar way.

The fine-cut distinctions in task performances observed in autism have been attributed to mindblindness and, in particular, blindness to the existence of representational mental states. A failure in autism to develop the normal capacity to represent the mind as a representational medium has been proposed to account for this mindblindness, which, in turn, explains the core triad of autistic symptoms: lack of pretend play, autistic aloneness, and abnormal communication (Baron-Cohen et al., 1993). A selective mindblindness of this type would explain the fine-cut distinctions listed above since (1) autistic individuals understand the distinction between physical representations (photographs) and reality but not the distinction between mental representations (beliefs) and reality; (2) visual perspective taking is intact in autism, and a visual percept cannot misrepresent the seen object in the same way that a belief can misrepresent the true state of affairs; and (3) autistic individuals understand basic emotions, and basic emotions, while consequent to events in the world (e.g., happiness due to a pleasant event), do not represent events in the world in the same way that beliefs do.

THEORIES OF AUTISTIC MINDBLINDNESS

Theoretical accounts of autistic mindblindness have been influenced by three prominent theories of how humans understand minds: theory-theory, modular theory, and simulation theory. Theory-theory is the view that a normal appreciation of minds depends on the acquisition of an everyday, or *folk*, theory about mental states, plus rules of inference concerning how mental states relate to behavior—for example, that people act intentionally to satisfy their desires in accord with their beliefs, even when those beliefs are false (see, e.g., Gopnik & Wellman, 1994). According to theory-theory, (1) normally developing children acquire a theory of mind in much the same way that they acquire other folk knowledge about the world; (2) experience is critical to achieve the necessary precursor stages to develop an adult theory of mind; and (3) no specialized ("modular") capacities are called upon when theory-of-mind knowledge is applied to predict and/or explain social behavior. In other words, we infer the contents of other people's minds by applying to the social domain the same *domain-general* inferential and problem-solving skills used in other areas of life. From a theory-theory perspective, autistic mindblindness might reflect the failure to acquire a normal theory of mental representations, possibly because autistic individuals fail to experience such critical precursor stages as imitative behavior. Alternatively, autistic mindblindness might reflect limitations of domain-general capacities, or limitations of performance (as opposed to competence). For example, impairments of hypothetical reasoning and inhibitory control might cause autistic individuals to be captured by the more salient objective facts of a situation (vs. other people's beliefs) or the literal (vs. figurative) meanings of other people's words; hence, they appear mindblind in the social domain. The latter account is often referred to as the executive-deficit account of autism, where *executive* refers to a set of higher-order domain-general capacities (e.g., working memory, strategic planning, and inhibitory control) required for complex and flexible problem solving.

Modular theory contrasts with theory-theory in proposing that there do exist specialized (i.e., *domain-specific*) capacities evolved for the specific purpose of mindreading (see, e.g., Baron-Cohen, 1995). Modular theory adopts a cognitive-science approach in order to model the representational structures that the human mind purportedly needs to compute and manipulate (online) when reasoning about observed behavior in terms of psychological causation. According to modular theory, what develops in a young child is not a "theory" of mental representations but a series of increasingly sophisticated cognitive mechanisms for representing agents. The endpoint is a theory-of-mind mechanism (ToMM; Leslie, 1994; Leslie & Roth, 1993) dedicated to inferring representational mental states by computing *meta*representations that link an agent to an aspect of reality via an *attitude* to the *truth* of a *proposition* (e.g., "Fred believes that *x* is true of *y*"). In other words, the ToMM is functionally specialized to represent *epistemic* mental states using what philosophers refer to as *propositional attitudes* (believing that, intending that, pretending that). The contrast here is with nonepistemic mental states such as percepts and emotions. The capacity to compute metarepresentations is deemed critical for decoupling representational mental states from reality, something that allows for *mis*representation. The ToMM interfaces with a (domain-general) selection processor (SP) responsible for inhibiting perceptions of current reality and selecting the appropriate counterfactual data to fill in the content of another person's mind.[2] From a modular ToMM perspective, autistic individuals might be blind to the existence of representational mental states, yet appreciate visual percepts and emotions, because the ToMM has not developed properly, whereas normally developing 2- to 3-year-olds might perform poorly on theory-of-mind *and* executive tasks, because the SP has not developed fully.

Simulation theory contrasts with both theory-theory and modular theory. According to simulation theory, people don't use a theory (of mind) very much of the time when appreciating other people's minds; nor do they appreciate other minds by computing metarepresentations in order to reason, in a third-person way, about observed behavior. Instead, mindreading is conceived of as a process of empathic role taking. In other words, in order to appreciate what another person will do (or think) in a particular situation, humans subjectively identify with that other person in order to run a simulation (in imagination) of being in the other person's *mental* shoes. From a simulation perspective, autistic mindblindness might reflect a general difficulty with simulating any hypothetical state of affairs (whether another person's belief or a set of future events being simulated in order to plan a strategic course of action), which then explains the co-occurrence of autistic asociality, lack of pretend play, and executive planning deficits observed in autism (Currie, 1995; Harris, 1993).

To summarize thus far, the psychopathological consequence of theory-of-mind impairment has been conceived of as autistic symptomatology and attributed to mindblindness. Theoretical accounts of autistic mindblindness have drawn upon (and reciprocally shaped) three prominent theories of how humans understand minds: theory-theory, modular theory, and simulation theory. However, autism is not the only form of psychopathology associated with theory-of-mind impairment; people with schizophrenia also perform poorly on theory-of-mind tasks.

THEORY OF MIND IN SCHIZOPHRENIA

Schizophrenic individuals demonstrate (1) a poor understanding of false belief and deception in stories, despite being matched to healthy controls on IQ; (2) a lack of appreciation of visual jokes when understanding the humor depends upon inferred mental states, despite being able to explain control jokes as well as healthy participants[3]; (3) a difficulty with sequencing picture-card stories that require inferences of false beliefs in order to determine the logical order of events, despite controlling for any difficulties with logical cause-and-effect reasoning in patients versus controls; and (4) an impaired capacity to go beyond the strict literal meanings of words in order to infer speakers' thoughts when speakers use either indirect hints or verbal irony, despite controlling for any limitations of verbal IQ and verbal memory in patients versus controls (for reviews, see Harrington, Siegert, & McClure, in press; Langdon, 2003). These findings present several challenges to current models of normal and abnormal mindreading that were developed, in part, to explain autistic mindblindness; the first of these challenges relates to the clinical phenomenology of schizophrenia.

Schizophrenia, unlike autism, is a *late-onset* neurodevelopment disorder with a typical onset in late adolescence to early adulthood. Approximately one in every 100 people is affected, males and females equally. Diagnosis is confirmed by the presence of any two or more of the following characteristic symptoms: delusions, hallucinations, disorganized speech, disorganized behavior, and negative symptoms. The latter include social withdrawal and apathy. Although some schizophrenic symptoms (e.g., social withdrawal) are similar to autistic symptoms, schizophrenia and autism are considered to be distinct clinical disorders based, in part, on epidemiological differences (e.g., age of onset, prevalence, and sex ratio). The clinical feature that diagnostically differentiates schizophrenia from autism is the presence versus absence of delusions (and hallucinations). Delusions are first-rank markers of schizophrenia and have been considered so since the 1950s. The most common delusion is persecution. Other delusions include grandiosity and loss of boundary (i.e., a breakdown in the normal sense of a barrier between self and others). Although the diagnostic criteria for schizophrenia list the presence of any two or more of the characteristic symptoms, the presence of delusions alone is sufficient to confirm diagnosis if these are of the bizarre type typically found in schizophrenia. The point here is that delusions (including persecutory delusions) are characteristic of schizophrenia but are not characteristic of autism; and a persecutory-deluded person with schizophrenia who believes that other people harbor hostile conspiratorial beliefs can hardly be thought of as someone who is blind to the existence of representational mental states.

Since schizophrenia is clinically heterogeneous, one might wonder whether it is only the nonparanoid patients who show theory-of-mind deficits. However, this idea can be ruled out; Harrington, Siegert, and McClure (in press) recently reviewed 25 studies of theory of mind in schizophrenia and concluded that the symptoms most consistently associated with theory-of-mind impairment in schizophrenia are persecutory delusions and disorganized speech. That people with schizophrenia are not mindblind was brought home to me recently when developing a joke appreciation task similar to that devised by Corcoran, Cahill, and Frith (1997). Like Corcoran and colleagues, I also observed patients who could explain control jokes as well as healthy participants, yet who failed to appreciate mental state jokes. For example, when asked to explain the cartoon illustrated in Figure 21.1, one patient responded, “Don’t get it. Maybe it’s the World Trade Center.” A second patient responded, “It’s a parody on the date being September 11. Someone in their apartment block is looking at the planes trying to kill some monster on the roof. It’s a ‘Thin Lizzie’ parody; they’re listening to ‘Killers on the Roof.’ ” And a third patient responded, “A guy steps out on the ledge to get the attention of the planes.” When this third patient’s attention was drawn to the giant fingers, he responded, “Maybe the monster’s throwing rocks on his head.”

I was not surprised by these responses; they were consistent with the findings of Corcoran and colleagues (1997). I was surprised, however, by the responses to the cartoon illustrated in Figure 21.2. I had initially intended to use this cartoon as a control joke since, although it depicted

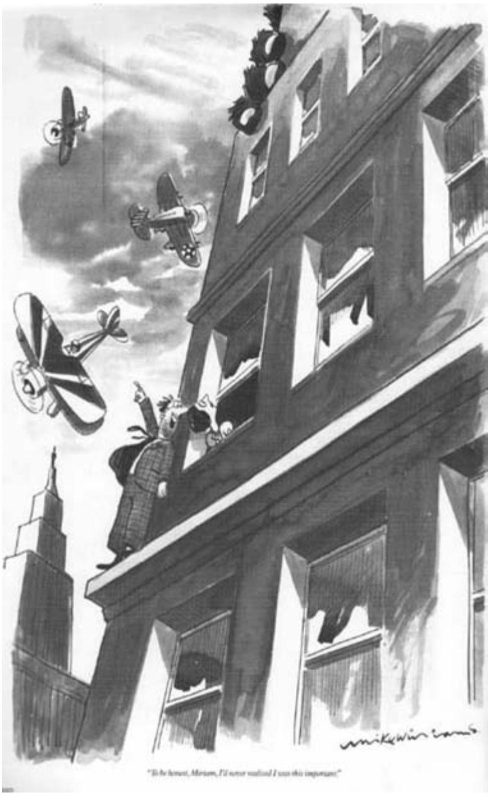
mikewilliams
"To be honest, Miriam, I'd never realized I was this important."

FIGURE 21.1. Mental state cartoon similar to those used by Corcoran, Cahill, and Frith (1997). Caption reads: "To be honest, Miriam, I never realized I was this important." Cartoon reproduced with permission of Punch Ltd.

shyness, all it required was an understanding of shyness as a behavioral disposition (not an intentional mental state)—for example, "He's begging for money, but he's shy and can't face people the way normal beggars would." I have since removed this cartoon from my battery since it prompted the following unexpected responses from the same three patients. The first patient responded: "There's a guy asking for money to help him overcome his acute shyness problem. He's obviously *a con-man trying to trick people* into feeling sorry for him." The second patient responded: "He's got his back to the street and *he's making out he's shy.* He's obviously got a problem." And the third patient responded: "I don't think he's got a shyness problem. He's *doing it to make people feel guilty* so he can collect money." These three patients (all with a history of persecutory delusions) seemed oblivious to what was going on in the

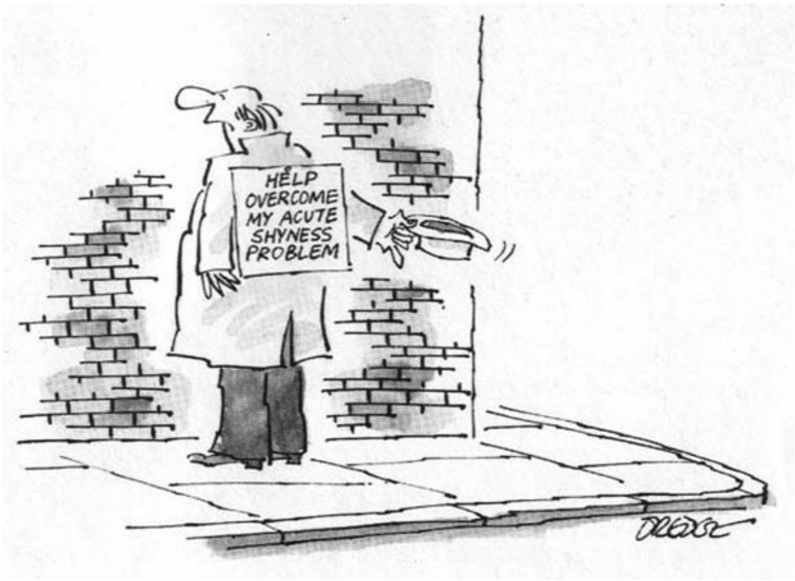

FIGURE 21.2. Cartoon initially intended as a control joke in a joke appreciation task styled on Corcoran, Cahill, and Frith (1997). Cartoon reproduced with permission of Punch Ltd.

minds of the cartoon characters depicted in Figure 21.1, yet inferred *intentional deception* on the part of the beggar in Figure 21.2.

Why are people with schizophrenia insensitive to the likely contents of other people's minds and prone to infer negative intentions where none exist? How do we explain theory-of-mind impairment in the absence of mindblindness in adults with schizophrenia?

That theory-of-mind impairment might reflect the failure, in childhood, to acquire a normal theory of minds (consistent with theory-theory) can be ruled out since schizophrenia is a late-onset disorder with a typical onset in late adolescence to early adulthood; if people with schizophrenia had failed, in childhood, to acquire a normal theory of minds, they would have come to the notice of clinicians long before the onset of their illness. Perhaps limitations of domain-general capacities cause people with schizophrenia to act as if they are mindblind in the social domain (an alternative idea still generally consistent with theory-theory). Making such a possibility plausible are the executive deficits that are common in schizophrenia (Morice & Delahunty, 1996), as well as the semantic deficits (i.e., degraded or disorganized conceptual knowledge; Goldberg & Weinberger, 2000) and poor attention and working memory (Nuechterlein, Edell, Norris, & Dawson, 1986). However, evidence suggests that theory-of-mind deficits in schizophrenia are not a secondary consequence of domain-general cognitive impairment, including poor conceptual reasoning (e.g., Doody, Gotz, Johnstone, Frith, & Owens, 1998) or executive deficits (Pickup & Frith, 2001). Harrington, Siegert, and McClure (in press) reported that 21 of the 25 schizophrenia studies that they had reviewed included control tasks to assess general intellectual ability, memory, or executive function; in all 21 cases, theory-of-mind deficits were found to be independent of control task performances. Similarly, Langdon, Coltheart, Ward, and Catts (2001a, 2002) found co-occurring executive deficits (poor executive planning and inhibitory control) and theory-of-mind deficits in two schizophrenia samples. In both cases, however, the executive deficits could not completely account for the theory-of-mind deficits; in particular, logistic regression analyses revealed that the theory-of-mind deficits and the executive deficits made significant independent contributions to discriminating the schizophrenic patients from the healthy controls. Rowe, Bullock, Polkey, and Morris (2001) reported similar results in patients with frontal brain damage. The implication here is that both types of deficit cannot reduce to precisely the same problem in all patients with schizophrenia or frontal brain damage and may instead reflect the disruption of neuroanatomically close, yet functionally dissociable, frontal regions. Consistent with this view, theory-of-mind deficits have been found in the complete absence of executive deficits in some patients with relatively circumscribed frontal damage (Bach, Happé, Fleminger, & Powell, 2000; Lough, Gregory, & Hodges, 2001).

Studies of theory of mind in schizophrenia suggest that something special, something that goes beyond general intellectual ability and executive function, is needed to appreciate other people's minds and that this something special is compromised in schizophrenia. Since it was modular theory (and not theory-theory) that had proposed the existence of domain-specific capacities for mindreading, might modular theory provide the more coherent account of theory-of-mind impairment in schizophrenia? Current modular theory, however, conceives of domain specificity in terms of a ToMM dedicated to representing epistemic mental states (i.e., beliefs and intentions) using the propositional attitudes. If a ToMM of this type were selectively disrupted in schizophrenia, one would expect patients to have difficulty with representing other people's beliefs and intentions (decoupled from reality); yet this cannot be the case since persecutory delusions are common in schizophrenia.

What of simulation theory? Here again we run into difficulties. This is because theory-of-mind impairment, as currently conceived from a simulation perspective, reflects a general difficulty with simulating any hypothetical state of affairs. However, such a general difficulty should impair equally theory-of-mind and executive capacities in schizophrenia; yet, we know that theory-of-mind deficits are independent of executive deficits in people with schizophrenia. Despite these reservations, a recent study of theory of mind and indirect speech comprehension in schizophrenia suggests that simulation theory might offer the better account of theory-of-mind impairment in people with schizophrenia.

SIMULATION THEORY AND SCHIZOPHRENIA

Langdon and colleagues (2002) were interested to see whether people with schizophrenia, like autistic individuals, show the co-occurrence of theory-of-mind deficits and poor appreciation of indirect speech. Happé (1993) had earlier investigated first-order and second-order theory of mind and comprehension of metaphors and irony in autism. First-order, in this context, refers to a capacity to infer "*A* believes that *x*," whereas second-order refers to a capacity to infer "*A* believes that *B* believes that *x*." The first-order versus second-order distinction purportedly helps to elucidate the distinction between the processes required for understanding of metaphors versus irony. More specifically, appreciation of indirect speech (whether metaphors or irony) requires, first, an understanding that speakers have thoughts that go beyond their words. Next, in order to understand a metaphor, a listener makes a first-order inference (e.g., in order to understand "My lawyer is a shark," a listener infers that the speaker thinks "My lawyer is a predator"). In contrast, in order to understand irony, a listener makes a second-order inference (e.g., in order to understand why a speaker says "What a fine friend" in the context of the supposed fine friend letting someone down badly, a listener infers that the speaker adopts an *attitude* of irony to an *expectation* conveyed via the words used; see Happé, 1993, for further discussion). In accord with this view, Happé found that performance deficits on first-order false-belief and deception tasks predicted poor appreciation of metaphors (and irony) in autism, whereas autistic individuals who succeeded on the first-order tasks (and only failed the second-order tasks) understood metaphors and failed to grasp only the irony. Findings of this type accord with disruption to a ToMM at different levels—for example, disruption at a first-order level will impair the capacity to pass first-order (and second-order) false-belief and deception tasks and to infer the thoughts of a speaker using metaphors (and irony). If, instead, the ToMM is only impaired at a second-order level, this will impair the capacity to pass second-order tasks and to appreciate irony but will leave intact the capacity to pass first-order tasks and to comprehend metaphors.

Langdon and colleagues (2002) found, however, that schizophrenic patients did not perform as the autistic individuals had. Performance deficits on a first-order false-belief task predicted patients' poor appreciation of irony but not poor understanding of metaphors. Some patients did find it difficult to understand metaphors; however, this difficulty was due, most likely, to independent semantic deficits. To clarify, the metaphor deficit and the first-order false-belief deficit made independent contributions to differentiating schizophrenic patients from healthy controls; hence, both deficits cannot reflect a common first-order theory-of-mind problem. Executive deficits were also ruled out as a full explanation of the metaphor problem (see Langdon et al., 2002, for details), suggesting that semantic deficits might be the cause. In accord with this view, Kintsch (2000) has proposed that metaphors are understood by activating semantic attributes of the metaphorical vehicle in order to find a contextual match (e.g., in order to understand "My lawyer is a shark," we activate "predator" features of the concept "shark"). If this is so, patients with schizophrenia who know that other people have minds and, hence, that speakers have thoughts that go beyond their words will be able to appreciate metaphors so long as their semantic capacities are intact; and some patients might fail to appreciate metaphors entirely due to disruption of their semantic (or conceptual) store.

The main point here is that the schizophrenic patients did not perform as the autistic individuals had. Hence, it seems unlikely that the same account used to explain the profile of task performances in autism (e.g., first-order vs. second-order disruption of the ToMM) will also explain the distinctive profile of task performances in schizophrenia. I suggest that the latter accords better with simulation theory. This is because a difficulty with simulating the subjective life of another person will impair a patient's ability to appreciate (1) the misguided thoughts of another person acting on a false belief and (2) the feelings of a speaker using irony, but it will not impair the capacity to appreciate metaphors. In contrast, metaphor comprehension will remain intact so long as the patient knows that there exist minds behind words (which I think that all people with schizophrenia do) and that the patient has no semantic deficits.

More direct support for a simulation account of theory-of-mind impairment in schizophrenia comes from a study of visual perspective taking (Langdon et al., 2001b). Whereas autistic individuals show a dissociation between impaired theory of mind and intact visual perspective taking (consistent with a ToMM model), people with schizophrenia do not. Langdon and colleagues tested visual perspective taking as comprehensively as possible since this capacity had not been tested elsewhere in schizophrenia. They used *item* and *appearance* questions (see below) and took account of limitations of visual working memory in schizophrenic patients. Reaction times (RTs) were also expressed as ratios of baseline RTs (on a visual matching task) since patients are generally slower than controls, regardless of the task. The same schizophrenic patients and healthy control participants who had earlier demonstrated selective theory-of-mind impairment in schizophrenia using a false-belief picture-sequencing task (Langdon et al., 2001a) took part.

It was hypothesized that, if the patients have difficulty with simulating other people's thoughts, these individuals should also have difficulty with simulating other visual experiences. We also suspected that difficulty with simulation might be more apparent on appearance (vs. item) questions. This is because item questions focus on the relative spatial locations of array features (see below) and, most likely, prompt geometric strategies (based on representing points in space and vectors), whereas appearance questions ask participants to imagine what it would *look like* to see an array from another perspective and, more likely, tap a capacity to simulate. Two instruction types were also used: *array rotation* and *viewer rotation* (see below). These were included to be comprehensive; we did not expect any differences here. Participants were presented with arrays of colored blocks on a stand; the stand was either fixed (for the viewer-rotation instructions) or a turntable (for the array-rotation instructions; see Figure 21.3). The questions were as follow:

- *Item question with array-rotation instructions*: "Imagine turning the stand so that the single dot is in front of you. Would the block in the FRONT on your RIGHT be BLUE?"
- *Appearance question with array-rotation instructions*: "Imagine turning the stand so that the single dot is in front of you. Would the blocks look like this?"
- *Item question with viewer-rotation instructions*: "Imagine moving to sit in the chair with the single dot. Would the block in the FRONT on your RIGHT be BLUE?"
- *Appearance question with viewer-rotation instructions*: "Imagine moving to sit in the chair with the single dot. Would the blocks look like this?"

For all appearance questions, a graphic image appeared on the screen (beneath the question) depicting an array of colored blocks, as if seen in perspective. Appearance questions varied in complexity: at the simplest level, three blocks were the same color and the fourth block a different color; at the most complex level, all four blocks were different colors.

I will focus on the results for the item questions and the simplest

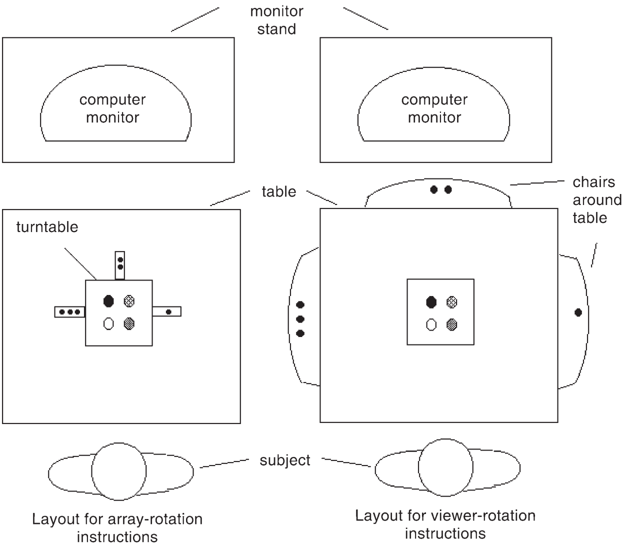

FIGURE 21.3. On the left, the layout for array-rotation instructions with lever arms extended to label the three sides of the turntable (one dot at 90°, two dots at 180°, three dots at 270°); on the right, the layout for viewer-rotation instructions with lever arms concealed and three labeled chairs placed around the table.

level of appearance questions since these questions were matched, as far as possible, for level of difficulty (i.e., in both cases, a participant could focus on a single critical block when responding). There were no significant differences in adjusted RTs. Patients and controls also judged item questions with equal accuracy: mean array-rotation accuracy was 91% for patients versus 95% for controls, and mean viewer-rotation accuracy was 98% for patients versus 97% for controls. In contrast, the patients made significantly more errors when judging the appearance questions, but only under the viewer-rotation instructions (59% for patients vs. 80% for controls), not under the array-rotation instructions (91% for patients vs. 86% for controls). This was a surprising set of results, not because of the difference between item and appearance questions (we had expected this) but because of the difference between array-rotation and viewer-rotation instructions.

Setting aside these differences for the moment, the first point to make here is that any evidence of co-occurring performance deficits on a theory-of-mind task and a visual perspective-taking task accords better with simulation theory (vs. a modular ToMM account). This is because disruption of a ToMM that is dedicated specifically to representing *epistemic* mental states should impair theory-of-mind task performances but should leave intact visual perspective taking. Schizophrenic patients did, however, show difficulty with visual perspective taking. At the same time, this difficulty occurred only with viewer-rotation instructions (not with array-rotation instructions), suggesting that schizophrenic patients do not have a general difficulty with simulation (if we take simulation to include the imagining of any hypothetical event).

When interpreting these clinical results and a set of similar results found when testing visual perspective taking in nonclinical adults with high levels of schizotypal personality traits (Langdon & Coltheart, 2001), Langdon and colleagues suggested that simulation for the specific purpose of mindreading might depend upon a capacity to use an *allocentric* (or world-centered) frame of reference in order to position self as only one subject among many, each apprehending a separate independent reality. Selective impairment of this domain-specific capacity (for intersubjective perspective taking) would then explain the poor appreciation of other people's beliefs and the difficulty with imaging other viewpoint-dependent perspectives in schizophrenia. In contrast, schizophrenic patients appear to have no difficulty with using an *egocentric* frame of reference in order to simulate the world shifting while they stay fixed as egocentric viewer (in order to judge appearance questions under array-rotation instructions). Nor do they demonstrate difficulty in using an allocentric *spatial* frame of reference in order to locate objects (rather than subjects) relative to a spatial world map (in order to judge item questions).

However, before proceeding further, it must be acknowledged here that my proposal of a domain-specific difficulty with intersubjective simulation in schizophrenia derives, in part, from the findings (reviewed earlier) of a link between poor appreciation of irony and theory-of-mind impairment in schizophrenia. I had interpreted these results in accord with simulation theory; however, there are many theories of irony. For example, the mismatch between the literal meaning of an ironical utterance and the conversational context might signal a denotative gap that the listener simply fills in by *negating* the literal meaning of the words used (e.g., "what a fine friend" becomes "what a terrible friend"), in which case there is no need to simulate in order to appreciate irony. As for the visual perspective-taking results, I cannot completely rule out a modular account conceived in terms of a ToMM dedicated to computing all metarepresentations (epistemic as well as perceptual). For example, schizophrenic patients might find visual perspective taking difficult because they cannot maintain a stable *representation* of an agent mentally *representing* the appearance of a 3–D object via a visual image (a form of perceptual metarepresentation). However, if patients were also found to have difficulty in appreciating other people's emotions, this would help strengthen the case that these individuals have difficulty with intersubjective simulation rather than difficulty with manipulating metarepresentations. This is because emotions, although decoupled from reality, do not stand in a *representational* relationship with external reality in the same way that beliefs and visual percepts do.

Langdon, Coltheart, and Ward (in press) recently investigated emotion attribution and theory of mind in schizophrenia. They used a false belief picture-sequencing task (used elsewhere to demonstrate selective theory-of-mind impairment in schizophrenia; Langdon et al., 2001a, 2002) to assess theory-of-mind abilities. This particular task was used because it provides appropriate control measures of general sequencing ability (see Langdon & Coltheart, 1999, for details). Emotion attribution was assessed using cartoon strips of events likely to elicit emotional reactions in story characters (see Figure 21.4). Characters' faces were blanked out, and cards depicting facial expressions were placed underneath cartoons in a prearranged incorrect order. Participants were instructed to imagine how the characters would be feeling in the depicted circumstances in order to match up the faces appropriately. Afterward, the participants were asked to name the emotions depicted in the faces. We found that the schizophrenic patients identified the emotions (from the cartoon faces) as well as the control participants, yet had greater difficulty in appreciating how story characters would be feeling in order to match emotions to circumstances appropriately. The patients also performed significantly more poorly on the false-belief picture-sequencing task; in neither case could poor IQ or general sequencing difficulties explain the patients' impairments. The implication here is that people with schizophrenia show evidence of a *modality-independent* difficulty with intersubjective perspective taking that impairs equally their appreciation of other beliefs, visual percepts, and emotions.

SUMMARY

The psychopathological consequence of theory-of-mind impairment has been traditionally conceived of as autistic symptomatology and attributed to mindblindness. Theoretical accounts of autistic mindblindness have drawn upon (and reciprocally shaped) three prominent theories of understanding minds: theory-theory, modular theory, and simulation theory. However, autism is not the only form of psychopathology associated with theory-of-mind impairment; people with schizophrenia, diagnostically differentiated from autistic individuals by the presence of delusions (including persecutory delusions), also show theory-of-mind

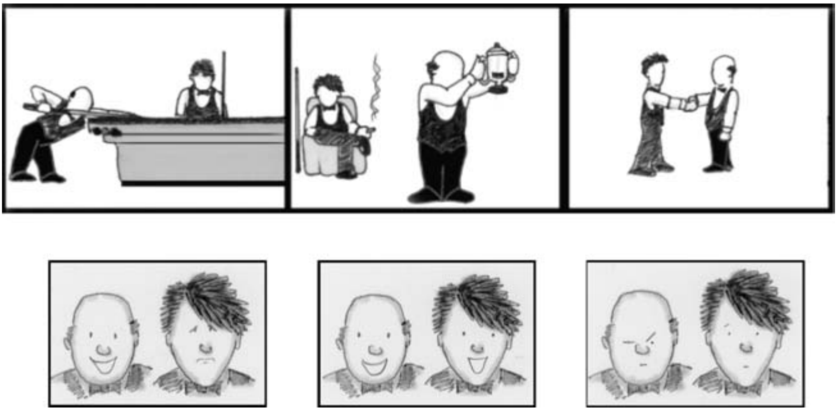

FIGURE 21.4. Cartoon strip from the emotion attribution task. Participants are instructed to imagine how each character is likely to be feeling in each panel and to rearrange the faces appropriately.

deficits. A schizophrenic patient with a persecutory delusion cannot be thought of as someone who is blind to the existence of representational mental states. This chapter has reviewed a body of work investigating the cause of theory-of-mind impairment in the absence of mindblindness in schizophrenia.

Since schizophrenia typically onsets in late adolescence to early adulthood, we can rule out a theory-theory account framed in terms of a failure, in childhood, to acquire a normal theory of minds. We can also rule out the idea that domain-general limitations cause patients to act as if they are mindblind in the social domain. This is because theory-of-mind deficits are independent of general intellectual impairment and executive deficits in schizophrenia. Evidence of selective theory-of-mind impairment in schizophrenia accords better with modular theory, the claim that domain-specific capacities that have evolved for the specific purpose of mindreading are selectively compromised in schizophrenia. However, current modular theory conceives of domain specificity in terms of a capacity to compute metarepresentations. This conception of domain specificity is unlikely to offer an adequate account of selective theory-of-mind impairment in schizophrenia for a number of reasons. First, difficulty in computing first-order metarepresentations should impair equally the capacities to infer first-order false beliefs and to appreciate metaphors; counter to that prediction, performance deficits on a first-order false-belief task in schizophrenia were associated with poor appreciation of irony (and not poor appreciation of metaphors). Second, a selective difficulty with computing metarepresentations using the propositional attitudes should impair false-belief inferences but should leave intact visual perspective taking; counter to that prediction, schizophrenic patients who showed performance deficits on a false-belief task also demonstrated difficulty in imaging other viewpoint-dependent visual perspectives. Third, domain specificity conceived in terms of metarepresentation (even without the propositional attitudes) is unlikely to prove adequate since schizophrenia patients not only have difficulty in appreciating other people's beliefs, they also show difficulty in appreciating other people's feelings and emotions, inner states that do not represent external reality in the same way that beliefs (or visual percepts) do.

I suggest, instead, a domain-specific difficulty with simulation (for the specific purpose of intersubjective perspective taking) as the underlying cause of theory-of-mind impairment (independent of general intellectual compromise and executive deficits) in people with schizophrenia. That is, these individuals have a selective difficulty with applying an allocentric frame of reference in order to map subjective experience relative to unique points of intersubjective space. Impairment of this domain-specific capacity then causes a difficulty with intersubjective perspective taking that affects equally the appreciation of other beliefs, percepts, and emotions in schizophrenia. Perhaps it is the functional capacity to map subjective experience relative to an allocentric frame of reference that allows us to be *both* in the shoes of another person (in imagination) *and* ourselves (in reality) at one and the same time while maintaining a mental separation between the real self and the simulated other (see Decety, Chapter 9, this volume, and Van Boven & Loewenstein, Chapter 18, this volume, for further discussion). If this functional capacity is impaired, as I suspect it is in people with schizophrenia, these individuals might occasionally lose themselves in the other, leading to loss of boundary experiences (Blakemore, 2003). Alternatively, they might retreat into their own solipsistic world, alienated from, and distrustful of, others and thus prone to persecutory thoughts.

ACKNOWLEDGMENTS

I am grateful to participants of *Other Minds: An Interdisciplinary Conference* held at the Institute of Cognitive and Decision Sciences, University of Oregon, September 2003, for helpful discussion of issues relevant to this chapter.

NOTES

1. A caveat here is that autistic individuals will fail to understand emotions when this understanding depends directly upon an ability to infer hidden false beliefs (e.g., in some cases of surprise).

2. The selection processor is similar to the inhibitory control component of executive function.
3. Control jokes depict *nonmental* humor—for example, a parachutist is depicted falling to earth beside a panda bear without a parachute, who looks across to the parachutist and says, “It’s no wonder we’re an endangered species really.”

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bach, L. J., Happé, F., Fleminger S., & Powell, J. (2000). Theory of mind: Independence of executive function and the role of the frontal cortex in acquired brain injury. *Cognitive Neuropsychiatry, 5*, 175–192.

Baron-Cohen, S. (1991). Do people with autism understand what causes emotion? *Child Development, 62*, 385–395.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Baron-Cohen, S., & Goodhart, F. (1994). The “seeing-leads-to-knowing” deficit in autism: The Pratt and Bryant probe. *British Journal of Developmental Psychology, 12*, 397–401.

Baron-Cohen, S., Spitz, A., & Cross, P. (1993). Do children with autism recognize surprise? A research note. *Cognition and Emotion, 7*, 507–516.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. J. (1993). *Understanding other minds: Perspectives from autism*. Oxford, UK: Oxford University Press.

Blakemore, S. J. (2003). Deluding the motor system. *Consciousness and Cognition, 12*, 647–655.

Charman, T., & Baron-Cohen, S. (1995). Understanding photos, models, and beliefs: A test of the modularity thesis of theory of mind. *Cognitive Development, 10*, 287–298.

Corcoran, R., Cahill, C., & Frith, C. D. (1997). The appreciation of visual jokes in people with schizophrenia: A study of mentalizing ability. *Schizophrenia Research, 24*, 319–327.

Currie, G. (1995). Imagination and simulation: Aesthetics meets cognitive science. In M. Davies & T. Stone (Eds.), *Mental simulation: Evaluations and applications* (pp. 151–169). Oxford, UK: Blackwell.

Doody, G. A., Gotz, M., Johnstone, E. C., Frith, C. D., & Owens, D. G. (1998). Theory of mind and psychoses. *Psychological Medicine, 28*, 397–405.

Goldberg, T. E., & Weinberger, D. R. (2000). Thought disorder in schizophrenia: A reappraisal of older formulations and overview of some recent studies. *Cognitive Neuropsychiatry, 5*, 1–20.

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. 257–294). Cambridge, UK: Cambridge Uniersity Press.

Happé, F. G. E. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition, 48*, 101–119.

Harrington, L., Siegert, R. J., & McClure, J. (in press). Theory of mind in schizophrenia: A critical review. *Cognitive Neuropsychiatry.*

Harris, P. (1993). Pretending and planning. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from autism* (pp. 228–245). Oxford, UK: Oxford University Press.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review, 7*, 257–266.

Langdon, R. (2003). Theory of mind and social dysfunction: Psychotic solipsism versus autistic asociality. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind: Implications for typical and atypical development* (pp. 241–270). Hove, UK: Psychology Press.

Langdon, R., & Coltheart, M. (1999). Mentalising, schizotypy, and schizophrenia. *Cognition, 71*, 43–71.

Langdon, R., & Coltheart, M. (2001). Visual perspective-taking and schizotypy: Evidence for a simulation-based account of mentalising in normal adults. *Cognition, 82*, 1–26.

Langdon, R., Coltheart, M., & Ward, P. B. (in press). Empathetic perspective-taking is impaired in schizophrenia: A study of emotion attribution and theory of mind. *Cognitive Neuropsychiatry.*

Langdon, R., Coltheart, M., Ward, P. B., & Catts, S. V. (2001a). Mentalising, executive planning, and disengagement in schizophrenia. *Cognitive Neuropsychiatry, 2*, 81–108.

Langdon, R., Coltheart, M., Ward, P. B., & Catts, S. V. (2001b). Visual and cognitive perspective-taking impairments in schizophrenia: A failure of allocentric simulation? *Cognitive Neuropsychiatry, 6*, 241–270.

Langdon, R., Coltheart, M., Ward, P. B., & Catts, S. V. (2002). Disturbed communication in schizophrenia: The role of poor pragmatics and poor theory-of-mind. *Psychological Medicine, 32*, 1273–1284.

Leekam, S. R., & Perner, J. (1991). Does the autistic child have a metarepresentational deficit? *Cognition, 40*, 203–218.

Leslie, A. M. (1994). ToMM, ToBY and agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). Cambridge, UK: Cambridge University Press.

Leslie, A. M., & Frith, U. (1988). Autistic children's understanding of seeing, knowing and believing. *British Journal of Developmental Psychology, 6*, 315–324.

Leslie, A. M., & Roth, D. (1993). What autism teaches us about metarepresentation. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from autism* (pp. 83–111). Oxford, UK: Oxford University Press.

Lough, S., Gregory, C., & Hodges, J. R. (2001). Dissociation of social cognition and executive function in frontal variant frontotemporal dementia. *Neurocase, 7*, 123–130.

Morice, R., & Delahunty, A. (1996). Frontal executive impairments in schizophrenia. *Schizophrenia Bulletin, 22*, 125–137.

Nuechterlein, K. H., Edell, W. S., Norris, M., & Dawson, M. E. (1986). Attentional vulnerability indicators, thought disorder, and negative symptoms. *Schizophrenia Bulletin, 12*, 408–426.

Pickup, G. J., & Frith, C. D. (2001). Theory of mind impairments in schizophrenia: Symptomatology, severity and specificity. *Psychological Medicine, 31*, 207–220.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526.

Reed, T. (1994). Performance of autistic and control subjects on three cognitive perspective-taking tasks. *Journal of Autism and Developmental Disorders, 24*, 53–66.

Reed, T., & Peterson, C. (1990). A comparative study of autistic subjects' performance at two levels of visual and cognitive perspective taking. *Journal of Autism and Developmental Disorders, 20*, *555–567*.

Rowe, A. D., Bullock, P. R., Polkey, C. E., & Morris, R. G. (2001). "Theory of mind" impairments and their relationship to executive functioning following frontal lobe excisions. *Brain, 124*, 600–616.

Tan, J., & Harris, P. L. (1991). Autistic children understand seeing and wanting. *Development and Psychopathology, 3*, 163–174.

Index

Page numbers followed by *f* indicate figure; *n*, endnote; and *t*, table

Actions
  as evidence of others' minds, 101
  inferences drawn from, 95*t*
  as visual cues in collaborative physical tasks, 94
Actor
  affect of
    perceiver impressions and, 161
    qualifying effects of, 162
  focus on, in attribution theories, 253–254
  self-perception of, 161
Affect
  actor
    perceiver impressions and, 161
    qualifying effects of, 162
  behavioral effects of, 170
Affective changes, after perspective taking, 48
Affective displays, observing, as mindreading strategy, 160–161
Agency, sense of
  brain areas involved in, 150
  empathy and, 153–154
Aggression, decreased, with perspective taking, 47
Alcoholism, shared experience of, empathy based on, 299, 302, 308*n*
Ambiguity
  referential, 259
  See also Trait terms, ambiguity of
Amnesia, infantile, episodic memory and, 200
Amygdala
  abnormalities in autism, 84
  in emotion recognition, 79
  in empathy, 152
  functions of, 80
  in perspective taking, 150
  stimulation by facial expressions, 148
Anterior paracingulate cortex, perspective taking and, 153
Anterior paracingulate/medial area of brain, in empathy, 153
Appearance–reality task, executive functioning in, 17–18
Ascriptions
  other. *See* Other-ascriptions
  self. *See* Self-ascriptions
Asperger's syndrome, empathy deficit in, 151
Atomism, conceptual, 64
Attribution theory
  covariation reasoning in, 107–108
  focus on actors versus situations, 253
  perspective taking and, 52
Augmenting, in social dynamics model of trait inference, 132
Autism
  brain abnormalities in, 84–86
  diagnostic criteria for, 324
  differential diagnosis of, 328
  executive deficit of, 326
  eye-gaze perception in, 83–84
  face recognition in, 85–86
  and failure to distinguish beliefs from reality, 324
  false-belief tasks and, 339*n*
  impairments in, 76–77
  joint attention in, 83–84

Autism (*continued*)
left hemisphere damage in, 192
mental verb deficits in, 217–218
mindblindness in, theories of, 325–327, 337–338
neurological abnormalities in, 83
processing of facial and emotional stimuli in, 84
raw behavior observation in, 28
versus schizophrenia, 337–338
symptoms of, 325
theory of mind in, 13, 323, 324–325
theory-of-mind deficits in, 217–218
training interventions in, 86
vocabulary acquisition and, 218
Autobiographical memory
development of, 200
ontogenesis of, 200
in self-ascriptions, 192, 194–195
self-metarepresentation and, 204

B

Barsalou's theory of perceptual symbol systems, 262
Behavior
accessing without mental state referencing, 27
imitative, as precursor to theory of mind, 325
as inputs/outputs of mindreading, 26–32
intentional versus accidental, in attribution theory, 253–254
intentional versus unintentional, infant ability to distinguish, 28
meaning of, context-dependence of, 256
mimicking, 36. *See also* Mimicry
misinterpretation of, self-predictions in, 293
observing, 40
as mindreading strategy, 160–161
raw observations of, 27–28
synchronization of, 36
terms for, ambiguity of, 259
transparent, 29–30, 31
in mindreading, 40
Behavioral changes, with perspective taking, 47
Body language
mindreading and, 27
See also Behavior; Facial expressions
Boundary ambiguity, 257–258
Boundedness, object, 257–258
Brain
abnormalities of, in autistic individuals, 84–86
social. *See* Social brain
social information processing in, 75–76, 86
and understanding of other minds, 75
Brain areas
in distinguishing self from nonself, 150
in empathy, 153–154
in episodic memory, 199
in eye-gaze perception, 78, 81
in facial expression observation/imitation, 147–148
left hemisphere
in autism, 192
in theory of mind, 192
in other-ascriptions, 202
in perception–action coupling, 147
in perspective taking, 149–150
right hemisphere
in schizophrenia, 192
in self-ascriptions, 192
in self-ascriptions, 202–203
in social processing, 80–81

C

Categorization, as late-developing activity, 69
Cheater detection, research on, 168
Children
episodic memory in, 200
folk psychology development in, 62–63
preschool. *See* Preschool years
Classification, as late-developing activity, 69
Cognition, embodied, 262
Cognitive development, research on, mindreading implications of, 272
Cognitive–affective mediating units, versus trait terms, 261
Collaborative physical tasks
challenges in, 93–96
common ground in, 94
conversational grounding in, 92
conversation's role in, 94
defined, 91–92
evidence about others' minds in, 100–102
and helper inferences about workers' minds, 97–98
situation awareness in, 92, 94
visual cues in, 91–105
empirical studies of, 96–103
and worker inferences about helpers' minds, 98–99
Common ground
in communication, 275
defined, 94
Common-coding theory, 146–147
Communication
and access to speaker's beliefs, 279
associative learning/priming in, 279

in autism, 325
common ground in, 275
face-to-face, tasks involved in, 223
functions of, 224
interaction and information aspects of, combined study of, 225–226
with language, 211
reference shifts in, 224
See also Conversation

Complement syntax, mastery of, 217

Concept acquisition
executive expression hypothesis and, 19–20
executive functioning in, 19–21

Concepts
activated, and activation of related concepts, 129–130
change in versus constancy of, 67
as classifiers, 68
folk psychological, 70*n*
independent existence of, 69
as independent versus private entities, 67–68

Conceptual alignment
as natural language feature, 241
during standardized interviews, 243–247
surveys of, 241–243
See also Conversation, conceptual alignment in

Conceptual atomism, 64

Conceptual deficit, in preschool theory-of-mind changes, 18–19

Conceptual development
denotational theory of, 65
holism in, 58–61, 63–67
modularity theory and, 62
paradox of, 56–72
protoconcepts in, 64–65
theory-theory and, 56–57, 62

Conceptual misalignment
awareness of, 248–249
defined, 240
importance of, 248
measurement of, 242

Conflict tasks, in executive functioning, 16

Consciousness, autonoetic/quasi-introspective, autobiographical memory in, 200

Constraint satisfaction
implicit theory of motive attributions and, 118–120
in social dynamics model of trait inference, 125, 132

Consumer Price Index Housing survey, conceptual alignment in, 243–245

Context
classification of, 28
defined, 28
in motive attributions, 110, 160–161
social information processing and, 86
trait meaning dependence on, 255
See also Situation

Conversation
in collaborative physical tasks, 93–94, 94
conceptual alignment in, 239–252
as natural feature, 241
questions about, 247–249
standardized surveys of, 241–243
studies of, 243–247
conceptual misalignment in, 240
diagnosis and correction of miscoordination in, 280
Gricean norms of, 93
interactional and informational aspects of, episodic representation of, 233–235
interlocutor's perspective during
study findings, 233–235
study methodology, 226–233
linguistic coordination in, 240–241
mental-state awareness and, 219
recall of, 229–232, 235
reference grounding in, 240
as route into other minds, 239
shared meanings in, 239–240
tracking in, 37
See also Communication; Social discourse

Conversational grounding
in collaborative physical tasks, 92
with visual cues, in collaborative physical tasks, 96–97

Covariation principle, implicit theory of motive attributions and, 116–118

Cross-cultural factors
in reliance on visual cues, 103
in theory of mind, 12, 17, 21

D

Deaf children, theory-of-mind and mental verb deficits of, 217–218

Definiteness, linguistic, 224
in face-to-face conversation, 233–234

Deixis, 93
in collaborative physical tasks, 95, 98–99
in conversation shifts, 224
in face-to-face conversation, 233–235

Delay tasks, in executive functioning, 16

Delusions, in differential diagnosis of schizophrenia, 328

Dementia, frontotemporal, 152

Denotational theory of conceptual content, 65

Dialogue. *See* Conversation

Discounting, in social dynamics model of trait inference, 132
Dispositional inference
  classic models of, 134–135
  explanatory coherence and goal-based knowledge structures and, 124–139
  multiple inference model of, 111
  situational restraints and, 134–135
  trait frames in, 129–136
Divorce, parental, empathy based on shared experience of, 299, 302–305, 308*n*
Dorsolateral prefrontal cortex, lesions of, empathy deficit due to, 151

E

Effect states, in mindreading, 40
Egocentrism
  adult, biased social judgments due to, 285
  in mindreading, 272–273, 277–278
  in perspective taking, 146
  recovery from, 280
  in social judgments, 293
Embodied cognition, 262
Emotion attribution, in schizophrenia, 337
Emotion sharing, perception–action coupling and, 146–148
Emotion–action complexes, 33–34
Emotional perspective taking
  behavioral misinterpretation in, 293
  dual judgment model of, 284, 285–287, 285*f*
  empathy gaps in, 284–297
  predicting others' reactions through, 286
  and prediction of feelings versus choices, 291–292
  self-predictions as basis for, 287–288
  in social interactions, 284
  in theory of mind, 284
  top-down versus bottom-up, 292–293
Emotional situations, predicting psychological impacts of, 291–292
Emotions
  contagion of, 36
  observed, hard-wired response to, 49
  orbitofrontal cortex in regulation of, 79
  others', perspective taking and, 45
  perception and generation of, 147–148
Empathic accuracy
  in close relationships, 310–322
    conventional wisdom about, 311, 320
    management implications, 311–312
  defined, 311
  in nonthreatening contexts, 312, 314
  relationships harmed by, 321
  relationships helped by, 321
  in relationship-threatening contexts, 314–316
Empathic accuracy model, 311–312, 313*f*
  assumptions of, 312
  basic interaction prediction in, 317–318
  negative effects of motivated accuracy in, 319–320
  positive effects of motivated inaccuracy in, 318–319
  preliminary tests of, 316–320
Empathic accuracy paradigm, Ickes's, 299–302
Empathic inaccuracy, motivated
  positive effects of, 311
  relationships helped by, 321
Empathic role taking, absence of, in autism, 326–327
Empathy
  amygdala in, 152
  as-if condition and, 145
  as complex social behavior, 153–154
  components of, 143
  deficits of
    in neurological disease, 152
    in neuropsychiatric disorders, 151–153
  enhancement programs for, 154
  executive functioning and, 144
  executive inhibition in, 148–149
  functional architecture of, 144*t*
  perception action mechanism in, 49
  perspective taking as source of, 143–157
  shared experience and
    future research on, 306–308
    study on, 299–306
  shared experience in, 298–309
  stimuli for, 144–145
  underlying mechanism of, 49
Empathy gaps
  impact bias and, 292
  in self and social predictions, 288–291
Empathy–altruism hypothesis, 149
Endowment effect
  basis for, 177
  duration of, 182, 186
  experiment on, 176–177, 176*t*
  thought experiment on, 181–182, 182*t*
Environment
  in collaborative physical tasks, 94
  inferences drawn from, 95*t*
  in perspective taking, 51
Executive emergence hypothesis, 22
  as complement to theory-theory and simulation theory, 22
  cross-cultural data and, 21
Executive expression hypothesis, concept formation and, 19–20
Executive functioning
  children's theories of minds and, 11–25
  in concept acquisition, 19–21

dimensions of, 16
empathy and, 144
impaired
in autism, 326
in schizophrenia, 330–331
inhibitory skill in, 17
measurement of, 13–14
in mental state understanding, 12
metarepresentation in, 15, 16
during preschool years, 13
theory of mind and, 14–19, 22
working memory in, 17
Eye gaze, inferences drawn from, 95*t*
Eye-gaze perception, 76–80
in assessing egocentrism in mindreading, 278
in autism, 83–86
behavioral evidence for, 77
brain areas involved in, 81–82
flexible use of, 82
isolable neural system for, 78
in mental state attribution, 81–82
modulation of, 79
neurological evidence for, 77–78
relation to other facial information processing, 79–80
as social module versus information processing, 78–79
theory of mind development and, 218
transition to joint attention, 82–83

F

Face recognition, by autistic individuals, 85–86
Facial expressions
inferences drawn from, 95*t*
mindreading and, 27
observation and imitation of, 147–148
relative importance of, 102–103
synchronization of, 36
as visual cues in collaborative physical tasks, 94–95
Facial information
processing of, 80–81
in autism, 84
eye-gaze perception in, 79–80
social and nonsocial processes in, 80–81
in social interactions, 80
False beliefs
past, recognition of, 198
self-ascription of, 193–194
False-belief tasks, 20, 194
for assessing egocentrism in mindreading, 277–278
for assessing theory of mind in schizophrenia, 337
in autistic individuals, 339*n*
children's ability in, 62
executive functioning in, 17–18
performance of
in autism, 332
factors influencing, 215
maternal discourse and, 214
pragmatic competence and, 218–219
semantic skills and, 215–217
in schizophrenia, 327
theory of mind and, 211
Faux pas task, 151
Folk psychology
concepts of, 70*n*
development of, 62–63
holism in, 60
inherent, domain-specific mechanisms in, 57
principles of, 56–57
Framing, traits as, 125–126
Frontopolar cortex
activation during perspective taking, 149
in empathy, 153
Frontotemporal dementia, 152
Fusiform face area
connectivity with amygdala, 79
in eye-gaze perception, 78
Fusiform gyrus
in autistic individuals, 84, 85
functions of, 81

G

Gaze. *See* Eye-gaze perception
Goals
inferences about, 124–125
target's, 50
in trait inferences, 127–129
See also Motives
Gricean norms of conversation, 93
Grounding
reference, 240
See also Conversational grounding
Group, mental state attributions for, 165

H

Holism, 70*n*
in conceptual development, 58–61, 63–67
in folk psychology, 60
in theory-theory, 58–61
Human associative memory model, 255
Huntington's disease, sympathy/empathy deficit in, 152

I

Ickes's empathic accuracy paradigm, 299–302
Imagination
  mental stimulation through, 175
  in perspective taking, 47
Impact bias, 291–292
Implicit theory, of motive attributions, 111, 116–120, 120
  constraint satisfaction and, 118–120
  covariation principle and, 116–118
Impression formation, theories of, trait invariance in, 255
Induction, evidence-based, 158–159
Infantile amnesia, episodic memory and, 200
Infants
  capacity to distinguish unintentional versus intentional behaviors, 28
  eye-gaze perception in. *See* Eye-gaze perception
  intentionality concept in, 58
  perspective taking in, 145
  social awareness of, 213–214
  transparent behaviors of, 30
Inferior parietal cortex, and distinguishing self from nonself, 150
Information
  conversation as source of, 223, 225–226
  perceiver's seeking of, 168–169
  preemptive sources of, 37–38
Inhibition
  in development of self-ascriptions, 191, 194
  executive, 17
    in regulation of self-perspective, 148–149
  impaired, in autism, 325
Intention attributions, affective displays in formation of, 161
Intentional agents, understanding people as, 82
Intentionality, interlocutor's, mental models of, 224–225
Intentionality concept
  development of, 28–29, 58
  as interpretational framework, 29
  mental states in, 29
  in mindreading, 39–40
Interactions
  evidence-based and extratarget strategies in, 165–168
  synchronization in, 36–37
Interlocutor's mind
  mental models of, 224–226, 234
  representation of, 223–238
Interpretability, presumption of, 249
Interviewing
  conceptual alignment in, studies of, 243–247
  standardized
    conceptual alignment in, 243–247
    term definition in, 242–243
Introspection, development of, self-ascriptions and, 191
Irony
  comprehension of, theory-of-mind skills and, 219
  versus metaphor, in schizophrenia, 332–333, 336
  theories of, 336–337
  understanding of, in autism, 332

J

Joint attention
  in autistic individuals, 85, 86
  defined, 82
  eye-gaze perception and, 76
  eye-gaze perception transition to, 82–83
  social rewards and, 84

L

Language, 209–222
  acquisition of, 212–214
  competence in, 211–212
  functions of, 211
  interdependence with theory of mind, 212–213
  intraindividual aspects of, 213, 215
    versus interindividual aspects of, 212
  mediating function of, 210
  pragmatics of, 218–219
  resolving ambiguity in, 275
  as route into other minds, 212–213, 213–219
  semantics of, 215–217
  social discourse context of, 214–215
  syntax of, 217–218
  theory of mind development and, 12, 213
Language processing, mental state attribution integration in, 280
Language use, in autistic individuals, 83
Language users, egocentrism of, 280
Learning, imitative, 82
Left hemisphere of brain. *See* Brain areas, left hemisphere
Limbic regions
  in autistic individuals, 85–86
  and distinguishing self from nonself, 150
Linguistic co-presence, 94
Linguistic coordination, conceptual misalignment masked by, 240

Listeners
conversation recall by, 229–232, 235
monitoring comprehension in, 275–276
perspective adjustment model and, 279
perspective taking by, 224–226
Logical inference, in perspective taking, 47, 48–49

M

Machiavellian intelligence hypothesis, 168
Medial prefrontal cortex, activation during perspective taking, 149
Memory
autobiographical. *See* Autobiographical memory
episodic, 195
in children, 200
defined, 199
human associative, 255
semantic, 195
defined, 199
working. *See* Working memory
Mental activities
after age 4, self-regulatory tasks in, 201–204
terms for, ambiguity of, 259
Mental concepts, trait concepts in, 254
Mental representations, during preschool years, 13
Mental rotation, 49
Mental state attributions
amygdala in, 152–153
in autistic individuals, 84–85, 85
bases for, 161
eye-gaze perception in, 76, 77, 81–82
group, 165
integration into language processing, 280
introspection in, 35
others' knowledge in, 37
stereotyping and, lack of research on, 163
theories of, 11–12
trait inferences and, 124
visual cues in, 95*t*. *See also* Visual cues
without behavior inputs, 30–32
Mental state certainty task, 20
Mental state concepts, formation of, 19
Mental state jokes, failure to appreciate, 328–330
Mental state reasoning
amygdala in, 152
executive functioning and, 12
pragmatic competence and, 218–219
in preschool years, 12–13
Mental state terms
acquisition of, 216
maternal use of, children's theory-of-mind skills and, 214–215
Mental states
awareness of, 211
as cause versus effect, 31
effects of, simulation theory and, 183
epistemic, disruption of, 335–336
inferences of, behavior and, 26–32
in intentionality concept, 29
knowledge of, and disambiguation of trait terms, 260–261
of others, capacity for knowing, 11
reasoning about, metarepresentation in, 216–217
representational, blindness to, in autism, 324–325
simulating features of, 176*t*, 182–183
terms for, 32
theory of, 56
Mental stimulation, 174–189
simulation theory and, 175
Mental verbs, 217
Mentalistic concepts, holistic meaning of, 60
Metacognition, development of, self-ascriptions and, 191
Metamind
development and tasks of, 202–203
self-regulatory work of, 203
Metamotor information, 203
Metaphor
experience-based, trait inferences and, 261–262
understanding of
in autism, 332
in schizophrenia, 333–334
Metarepresentation
and conflict and delay tasks, 16
development of, 214, 219
in executive functioning, 15, 16
in mindreading, 273–275
and reasoning about mental states, 216–217
self, theory of mind and, 202–203
in self- and other-ascriptions, 200–201
in theory of mind, 15, 16, 197
Metonymy, for scripted activities/mental events versus objects, 258
Mimicry
behavior, 36
in perspective taking, 47, 48, 49
Mind, 32–36
development of, 75
interlocutor's. *See* Interlocutor's mind
others.' *See* Other minds
as set of subsystems, 76
theory of. *See* Theory of mind
Mindblindness
autistic, theories of, 324–327, 337–338
in differential diagnosis of schizophrenia, 328
theory-of-mind impairment and, 337

Mindreading, 26–55
activities described as, 273
allocentric versus egocentric framework in, 336, 339
behavior as inputs and outputs of, 26–32
brain representations and, 75
defined, 26
developmental course of, 272
egocentrism in, 272–273, 277, 278
correction of, 273
evidence-based strategies in, 165–168
evolution of, 26, 63
extratarget strategies in, 165–168
facile, 39
as function of social brain, 75–76
as high versus low-level activity, 36–39
metarepresentations in, 273–275
nonbehavioral sources of, 163–165
nonnormative/primate research and, 272
in normal adults, 271–283
psychological tools in, 39
reasoning by analogy in, 170
and research on cognitive development, 272
strategies for, 158–173, 159*f*
evidence-based, 160–162
evidence-based versus extratarget, 169–170
evidence-based–extratarget interaction, 165–168
extratarget, 163–165
negative social intention and, 170
projection, 159
stereotyping, 159
which-tool-when contingencies, 160
as subsystem of mind, 76
underlying processes in, 39–40
Modular theory, 12
autistic mindblindness in, 326
cognitive skills in, 212
concept development and, 62
executive emergence theory and, 22
schizophrenia in, 331, 338
self-ascription in, 198
Motherhood, shared experience of, empathy based on, 299, 302–304, 308*n*
Motivational changes, with perspective taking, 47
Motive attributions, 106–123
constraint satisfaction and, 118–120
context of, 109, 110
covariation principle and, 116–118
impression formation process and, 109–110
intentionality of, 108–109
research on, 106
situations in, 160–161
speed of, 109
spontaneity of, 108–109, 111
theories of, 111–120
implicit, 111, 116–120, 120
simulation, 111–116, 120
trait attributions and, 110
Motives
importance of, 107–108
target's, 50
See also Goals
Motor behavior, in autistic individuals, 83
Motor deficits, in autistic individuals, 85
MT area of brain, 78, 80

N

Narrative development, theory-of-mind skills and, 219
Narratives, traits as, 126
Neuropsychiatric disorders, empathy deficits in, 151–153
Newborn, eye-gaze perception in, 81

O

Objects
boundedness of, 257–258
metonymy for, 258
Observer
accuracy of judgments of, 52
as information source in perspective taking, 46, 51
relationship with target, 50
See also Perceiver
Observer–target merging, 47
Orbitofrontal cortex
in emotion regulation, 79
functions of, 80–81, 84
lesions of, empathy deficit due to, 151
Other minds
conversation as route into, 239
everyday solutions to problem of, 158–173. *See also* Mindreading, strategies for
language as route into, 212–213, 213–219
understanding, social discourse in, 214–215
visual cues as evidence of, 91–105. *See also* Visual cues
Other-ascriptions
brain areas involved in, 202
defined, 191
development of, versus self-ascription development, 198, 204
egocentricity of, 196–197
metarepresentation in, 200–201

P

Perceiver
- forgivingness in, 163
- information sought by, 168–169
- projection relinquished by, 167–168
- stereotyping relinquished by, 167
- *See also* Observer

Perception–action coupling
- brain areas involved in, 147
- to emotion sharing, 146–148
- in empathy, 49

Perceptual point of view, target's, 49
Perceptual symbol systems, Barsalou's theory of, 262
Persecution delusions, in differential diagnosis of schizophrenia, 328
Personality, five-factor model of, 255–256
Personality traits. *See* Traits
Perspective adjustment model, mindreading and, 279
Perspective taking
- accuracy as goal in, 52
- changing research emphases in, 52–53
- constituents of, 44–55
  - model for, 45*f*
  - natural, 48–50
  - neglected, 50–51
- development of, 145, 211
- egocentricity in, 146
- emotional. *See* Emotional perspective taking
- in empathy, 153
- environmental information in, 46, 51
- functional neuroimaging of, 149–150
- goals of, lack of research on, 52–53
- information sources in, 46
- by listeners during conversation, 224–226
- mental flexibility in, 148–149
- observer as information source in, 51
- per se, as research focus, 53
- processes in, 46–47
- purpose of, 45–46
- results of, 47–48
- in schizophrenia, 339
- self, as default mode, 145–146
- as source of empathy, 143–157
- stereotyping and, 51
- target of, 46
- unintended effects of, 52–53
- visual
  - in autism, 325
  - in schizophrenia, 335–336

Physical co-presence, 94
Physical processes, simulation versus theory of, 176
Polysemy, of trait terms, 256, 257
Posterior temporal lobe, in eye-gaze perception, 78
Pragmatics
- competence in, theory of mind and, 218–219
- and interrelation of language and theory of mind, 213

Prefrontal cortex
- damage to, Stroop task errors and, 280
- in empathy, 153
- lesions of, empathy deficit due to, 151
- metamind development in, 202
- in sense of agency, 154

Preschool years
- executive functioning in, 13
- mental representation capacity in, 13
- mental state understanding in, 12–13
- theory-of-mind changes in, 18–19, 22

Presumption of interpretability, 249
Pretense-driven offline simulation, 175–176
Primates
- eye-gaze perception in, 77, 81
- research on, human mindreading implications of, 272

Projection
- as default starting point, 166
- as mindreading strategy, 159, 163–165
- perceived versus actual similarity in, 164, 170
- in perspective taking, 47
- versus stereotyping, 164
- in trait inferences, 134

Propositional belief, understanding of, 62
Protoconcepts, 64–65
Proto-declarative pointing, 82
Prototypes, generic, 171*n*
Psychology, folk. *See* Folk psychology

R

Reasoning by analogy, in mindreading, 170*n*
References, grounding, 240
Relationships
- danger zones for, empathic accuracy and, 312
- empathic accuracy/inaccuracy in. *See* Empathic accuracy; Empathic inaccuracy
- observer–target, 50

Representation, through language, 211
Representingness
- defined, 190
- self. *See* Self-representingness

Right hemisphere of brain. *See* Brain areas, right hemisphere
Right inferior parietal lobule, in sense of agency, 154

S

Sally–Anne task, 323–324
Schemas, traits as, 125–126
Schizophrenia
  versus autism, 337–338
  diagnostic symptoms of, 327–328
  differential diagnosis of, 328
  emotion attribution in, 337
  executive deficits in, 330–331
  modular theory and, 338
  onset of, 338
  right hemisphere damage in, 192
  simulation theory and, 331–337, 339
  theory of mind in, 323–342
  theory-of-mind deficits in, 337–338
  theory-theory and, 338
Self-ascriptions
  autobiographical memory in, 192, 194–195
  brain areas involved in, 202–203
  cognitive asymmetry in, 194
  components of, 192–193
  defined, 190–191
  development of, 190–206
    versus other-ascription development, 198, 204
    self-representingness and, 192
    visual evidence in, 195, 196
  developmental asymmetry in, 191–192
  executive symmetry in, 194
  experiments on, 194–195
  false beliefs and, 193–194
  inhibition and, 191, 194
  introspection and, 191
  memories of past attitudes in, 199–201
  metacognition and, 191
  metarepresentation in, 200–201
  in modular theory, 198
  of past false belief, 193–194
  representingness-sensitive, 200
  right hemisphere in, 192
  self-representingness and, 192
  in simulation theory, 197–198
  theories of, 195–199
  in theory-theory, 195–196
  thought identification and, 191
  See also Self-representingness
Self-metarepresentation
  autobiographical memory and, 204
  brain areas involved in, 202–203
  theory of mind and, 202–203, 204
Self-perceptions, of actor, 161
Self-perspective
  as default mode, 145–146
  regulation of, 148–149
  See also Egocentrism
Self-predictions
  as basis for emotional perspective taking, 287–288
  empathy gaps in, 288–291
Self-representingness
  development of, 197, 203
  and development of self-ascriptions, 192
  elusiveness of, 192–195
  separate sense of, 199
  See also Self-ascriptions
Semantic skills, performance on false-belief tasks and, 215–216
Semantics, defined, 215
Shame, observations of, positive impressions due to, 161
Simulation
  in perspective taking, 47
  pretense-driven offline, 175–176
Simulation theory, 12
  autistic mindblindness in, 326–327
  constraining, 178–183
  empirical implications of, 183, 185
  empirical testing of, 186–187
  evidence supporting, 183
  executive emergence hypothesis as complement to, 22
  focus of, 174
  mental causation in, 179–182
  mental state features in, 182–183
  mental stimulation of mind and, 175
  on motive attributions, 111–116, 120
  offline mental states in, 177–178
  origins of, 174
  schizophrenia in, 331–337, 339
  self-ascriptions and, 197–198
  situation in, 179
  versus theory of physical processes, 176
  versus theory-theory, 184–187
Situation
  juxtaposed, predictions for, 184–185
  mental stimulation of, 179
  neglect of
    in attribution theories, 253–254
    in personality theory, 254
  physical effects of, simulation theory and, 183
  in simulation theory, 179
  See also Context
Situation awareness, in collaborative physical tasks, 92, 94
Smiling, EEG activity during, 147
Social brain, 75–90
  contradictions of concept, 76
  eye-gaze perception and, 76–80. *See also* Eye-gaze perception
  model for, 75–76
  theories supporting, 78–79

Social cognition, deficits of, in neurological disease, 152
Social discourse
  influences of, on understanding other minds, 214–215
  as route into other minds, 213
  See also Conversation
Social dynamics model of trait inference
  actors, objects, actions in, 131
  applications of, 133–136
  assumptions of, 125
  constraint satisfaction in, 125, 132
  example of, 130–133
  meaning level of, 131–133
  organization of, 129, 130*f*
  scenarios in, 131
  social knowledge structures in, 125
Social information, brain's processing of, 86
Social interactions
  in autistic individuals, 83
  facial information in, 80
Social knowledge structures, goal-based, in social dynamics model of trait inferences, 125
Social psychology, personality trait meaning in, 255–256
Social reasoning, egocentric bias in, 274
Social referencing, 82
Source-of-knowledge tasks, 20
Speakers, access to beliefs of, successful communication and, 279
Speech acts, epistemic modality of, 230
Speech–act grammar, in face-to-face conversation, 233–234
Stereotyping
  decreasing potency of, 167
  as default starting point, 166
  and meaning of trait terms, 259
  mental state inferences and, lack of research on, 163
  as mindreading strategy, 159, 163–165
  negative, resistance of, 169
  perceived versus actual similarity in, 164, 170
  perspective taking and, 51
  versus projection, 164
  reduced application of, in perspective taking, 47
  in trait inferences, 134
Stroke, empathy deficit in, 151
Stroop task, prefrontal damage and, 280
Substances, reidentifying, 68–69
Superior temporal sulcus
  abnormalities in autism, 84
  as evidence of social brain, 78–79
  in eye-gaze perception, 78, 79
  functions of, 80
  stimulation by facial expressions, 147–148
Suspect theory, 59
Syntactic skills, false-belief task performance and, 217–218
Syntactic theory, executive emergence theory and, 22
Syntax, complement, mastery of, 217

T

Target
  accurate judgment of, 47
  affective displays of, in shaping inferences, 161
  goals of, 50
  motives of, 50
  perceptual point of view of, 49
  of perspective taking, 46, 48
  relationship with observer, 50
  sensory experience of, 49
  stereotyping of, 51
Tasks. *See* Collaborative physical tasks; specific tasks
Telecommunications, collaborative physical tasks through, 96
Thalamus, and distinguishing self from nonself, 150
Theory of mind
  absence of processing model of, 274–275
  acquisition of, 325
    language and, 210
  approaches to, 11–12
  in autism, 13, 324–325
    versus schizophrenia, 323
  breakdown of, with frontotemporal dementia, 152
  cognitive–affective mediating units in, 261
  cultural variation and, 12
  definitions of, 210–211, 323
  development of, 62
    language development and, 213
    linguistic environment and, 216
  domain-specific and -general abilities in, 21–22
  early literature on, absence of representingness in, 193
  emotional perspective taking in, 284
  evolution of, negative social intention and, 168–169
  executive functioning and, 12–14, 17–19
  eye-gaze perception and, 218
  as factor in executive functioning development, 14–17
  folk psychology and, 57
  in human interactions, 271
  impaired
    mindblindness and, 337
    in schizophrenia, 330–331, 332, 337–338

Theory of mind (*continued*)
 inherent nature of, 60, 63
 interdependence with language development, 212–213
 language development and, 12
 left hemisphere in, 192
 metarepresentation in, 15, 16
 neural substrates of, 274
 pragmatics competence and, 218–219
 representational, 187*n*
 in schizophrenia, 323–342
 self-metarepresentation and, 204
 tasks of, executive functioning in, 17–18
 terminology of, 70*n*
 trait concepts in, 254
 in trait inferences, 261
Theory of mind mechanism
 in autism, 326
 in schizophrenia, 331, 332–333, 335–336
Theory-theory, 11–12
 autistic mindblindness in, 325–326
 concept development and, 62
 conceptual grounding of, 56–57
 dilemmas of, 61–62
 executive emergence hypothesis as complement to, 22
 holism in, 58–61
 schizophrenia in, 338
 self-ascriptions in, 195–196
 versus simulation theory, 186–187
 testing for use of, 184–186
Thinking, re-creative, autobiographical memory in, 200
Thought identification, development of, self-ascriptions and, 191
Thoughts, others'
 perspective taking and, 45
 See also Other minds
Three Mountains Task, 49
Tobacco Use Supplement to the Current Population Survey, conceptual alignment in, 245–247
Tracking. *See also* Eye-gaze perception
 conversational, 37
Trait frames, in dispositional inference, 129–136
Trait inferences
 causal attributions in, 110
 development of, 125. *See also* Social dynamics model of trait inference
 experience-based metaphor and, 261–262
 mental state attributions and, 124
 motive inferences and, 120
 projection versus stereotyping in, 134
 social dynamics model of. *See* Social dynamics model of trait inference
 theory of mind in, 261
 unconscious versus conscious processes in, 254
Trait terms
 ambiguity of, 253–267
  boundary, 257–258
  homonymous versus polysemous, 258–259
 applications and characteristics of, 259–260
 disambiguation of, 260–262
 versus goal descriptions, in impression formation, 260
 multiple versus single meanings of, 256
 polysemy of, 257, 261
 stereotyping and, 259
 views of, 254–257
Traits
 Big Five global factors of, 255
 as frame-based structures, 125–126
 as goal-based, 126, 127–129
 as invariant versus context-dependent, 255–256
 as narratives, 126
Transparent behaviors, 29–30, 31
 in mindreading, 40

V

Ventromedial prefrontal cortex
 activation during perspective taking, 149
 in empathy, 153
 lesions of, empathy deficit due to, 151
Video technologies, conversational grounding with, 97
Visual attention deficits, in autistic individuals, 85
Visual cues, 91–105
 in collaborative physical tasks, 93–96
  empirical studies of, 96–102
 decompositional approach to, 95–96
 hierarchy of, 100–101
 importance of, cultural differences in, 103
 relative importance of, 102–103
 roles of, 92
 sources of, 94, 95–96
 uses of, 102

W

Wiggins's Interpersonal Circumplex, 126
Working memory
 in executive functioning, 17
 in mental state concept formation, 19
 visual, in schizophrenia, 334