

OXFORD

UNDERSTANDING OTHER MINDS

PERSPECTIVES FROM DEVELOPMENTAL
SOCIAL NEUROSCIENCE

EDITED BY SIMON BARON-COHEN,
HELEN TAGER-FLUSBERG AND
MICHAEL V. LOMBARDO



Understanding Other Minds

This page intentionally left blank

Understanding Other Minds

Perspectives from
Developmental Social
Neuroscience

Edited by

Simon Baron-Cohen

Helen Tager-Flusberg

Michael V. Lombardo

OXFORD
UNIVERSITY PRESS

OXFORD

UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© Oxford University Press 2013

Author's contribution to the Work was done as part of the Author's official duties
as a NIH employee and is a Work of the United States Government. Therefore,
copyright may not be established in the United States (Chapter 20).

The moral rights of the authors have been asserted

First Edition published in 1993
Second Edition published in 2000
Third Edition published in 2013

Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2013936562

ISBN 978-0-19-969297-2

Printed and bound by
Ashford Colour Press Ltd, Gosport, Hampshire

Oxford University Press makes no representation, express or implied, that the
drug dosages in this book are correct. Readers must therefore always check
the product information and clinical procedures with the most up-to-date
published product information and data sheets provided by the manufacturers
and the most recent codes of conduct and safety regulations. The authors and
the publishers do not accept responsibility or legal liability for any errors in the
text or for the misuse or misapplication of material in this work. Except where
otherwise stated, drug dosages and recommendations are for the non-pregnant
adult who is not breastfeeding.

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

Preface

Understanding Other Minds: What's new?

Simon Baron-Cohen, Helen Tager-Flusberg, and
Michael V. Lombardo

As editors of *Understanding Other Minds* (3rd edn, henceforth *UOM-3*), we are proud to have compiled such an exciting set of new chapters, by such an internationally impressive set of scholars, addressing what some regard to be the central psychological process separating humans from all other animals: namely, the ability to imagine the thoughts and feelings of others, and to reflect on the contents of our own minds. This drive and capacity to attribute mental states to others has for 30 years gone under the rubric of possessing a “theory of mind” (ToM).

In *UOM-1* (1993) and *UOM-2* (2000), we brought together the state of the art in research into ToM during each decade, bringing together scientists and philosophers from fields as diverse as developmental psychology, psychiatry, and clinical psychology, neuroscience, primatology, and philosophy. The aim was to understand the nature of ToM by studying its development, its impairment, its brain basis, its evolution, and its theoretical baggage. For those volumes, we were joined by a third editor, Donald Cohen, who tragically passed away far too young, after a difficult battle with cancer, and who brought a psychiatrist's perspective to bear to this fundamental field.

The need for a new edition of this book comes about because the field has not stood still over the past decade—on the contrary, the field has continued to attract some of the best minds in the effort to understand our mind. So what's new in *UOM-3*? First, we have a new co-editor, Mike Lombardo, who is an example of how the field has blossomed via a new generation of talented young developmental social neuroscientists interested in ToM both from the standpoint of typical development and its atypical expression in conditions such as autism. Secondly, as our understanding of the biology underlying ToM has deepened, so has our understanding of its development, cross-cultural expression, and its atypical role in a variety of neurodevelopmental conditions. In this Preface we provide a brief summary of what a reader of *UOM-3* can expect, reflecting these new developments in the field.

Development

Victoria Southgate (Chapter 1), who opens this volume at the earliest stages of development, reveals how infancy research demonstrates that ToM is present much earlier than previously believed. She reviews exciting work from her own and other laboratories, suggesting that

infants expect others' behavior to be congruent with their own beliefs, even in the first two years of life. She argues that infants' performance on the tasks used to tap these abilities do not just reflect behavior-reading, but actually reflect mindreading, even at this young age. Andrew Meltzoff and Alison Gopnik (Chapter 2) review training studies suggesting that a ToM "module" does not just "turn on" in the child, but that development is influenced by experience, evidence, and learning. Furthermore, they argue that the child's initial state contains Bayesian "priors" that constrain learning, one example being the principle that other minds are "like-me". They describe elegant new experiments with young children to support their ideas, for example, from understanding other people's visual perspectives, and conclude that children's ToM is plastic enough to accommodate to the specific culture in which they find themselves.

Johannes Roessler and Josef Perner (Chapter 3) address a classic question from ToM research (why 3-year-olds typically fail false belief (FB) tasks), by arguing that 3-year-olds are "teleologists". By dissecting what young children think other's "ought" to do in a situation, Roessler and Perner offer an explanation for why young children's explicit ToM (where they make errors) is at odds with their own implicit ToM (which, as Southgate shows, they already possess).

Ian Apperly (Chapter 5) reminds us that ToM development doesn't stop in childhood, and that by studying ToM in adults we see that some forms of ToM require effort, whilst others are effortless and even automatic. He makes a claim for the existence of two systems, and links this to the infancy work and to the neural basis of ToM. Later in this volume, Alvin Goldman and Lucy Jordan (Chapter 25), from their perspective as philosophers, update the cognitive debate between "simulation theory" and "theory theory" as mechanisms underlying the development of a ToM.

Cross-cultural perspectives

Henry Wellman and Candida Peterson (Chapter 4) provide a striking graph showing that across eight different cultures, the same transition is seen between approximately 3- and 4-year-olds (with some cultural variation in ages, but not in trajectory) in passing FB tests. They also report their efforts to create a ToM scale that can be used not only across cultures, but also across medical conditions, and describe their investigations into how ToM develops differently and later in those deprived of hearing spoken language (deaf children). Their results sit comfortably with Meltzoff and Gopnik's conclusions that the nature of the input a child receives affects the way in which ToM develops. Liane Young and Adam Waytz (Chapter 6) go one step further, to explore the interesting claim that we use our ToM most when we make moral judgments. David Kenny (Chapter 7) guides us through the array of standardized measures that exist to study "judgment accuracy", a factor within "emotional intelligence", which overlaps with ToM, reminding us of the importance of psychometric issues in how we measure ToM.

Electrophysiology and functional neuroimaging

Mark Sabbagh (Chapter 8) picks up the theme of the neural basis of ToM by discussing encephalographic recordings (EEG)/event-related potential (ERP), making a claim for the N270 playing a key role, and discussing mu-suppression during both intentional action and perception of intentional action. Jorie Koster-Hale and Rebecca Saxe (Chapter 9) give us a tour of the functional magnetic resonance imaging (fMRI) literature on ToM, reminding us that, whereas in *UOM-2* only four studies were reviewed, today there are over 400! Right temporo-parietal junction (RTPJ) and ventromedial prefrontal cortex (vMPFC) are, they argue, well-replicated

ToM regions and “one of the most remarkable scientific contributions of human neuroimaging, and the one least foreshadowed by a century of animal neuroscience”. These regions, they argue, do not work in isolation, but are part of a network. Despite the widely differing experimental paradigms different investigators have employed, consistently similar brain regions are activated.

Neurological lesion studies

Dana Samson and Caroline Michel (Chapter 10) update our knowledge about ToM from studies of brain damage. They describe patient WBA who, following a stroke and acquired damage to his right lateral prefrontal cortex, suffers from an inability to set aside his own perspective. A second patient, PH, following a left-hemisphere stroke, suffered from an inability to process grammar, but his ability to pass false belief tasks remained unaffected. They argue that this suggests that once ToM is established, syntactic ability plays a minor—if any—role. A third patient, CM, with semantic dementia and atrophy of the left temporal pole, struggled to understand mental state words, but had no difficulty understanding others’ intentions on non-verbal tasks. These valuable “natural experiments” enable “fine cuts” in the neuropsychology of ToM.

The neural basis of empathy

Anat Perry and Simone Shamay-Tsoory (Chapter 11) extend this approach to the study of empathy, fractionating it into “emotional” and “cognitive” empathy, making a case from both lesion and fMRI studies for inferior frontal gyrus (IFG) being central to emotional empathy, anterior cingulate cortex (ACC) and insula being central to pain perception, with each of these linking to the amygdala. In contrast, they present the evidence for cognitive empathy being a circuit comprising TPJ, superior temporal sulcus (STS), vmPFC/orbito-frontal cortex (OFC), dorsolateral (dlPFC) and dorsomedial prefrontal cortex (dmPFC). They also look at evidence from fMRI studies to show how these regions overlap and differ in neuropsychiatric conditions, such as autism, schizophrenia, and psychopathy, and how the two components of empathy are both independent and yet interact.

Cade McCall and Tania Singer (Chapter 12) also consider the brain bases of empathy, delineating the “pain matrix” through experiments. An example is where the observer sees a Q-tip stroking a hand or a needle puncturing a hand, which gave rise to the discovery that parts of this “matrix” are active when we experience pain and when we observe another person in pain, validating a “mirror system”. Jamil Zaki and Kevin Ochsner (Chapter 13) pick out an Experience Sharing System (ESS), distinct from a Mental State Attribution System (MSAS), as what they call a “tale of two systems”. This converges on the emotional vs. cognitive empathy systems delineated by Perry and Shamay-Tsoory.

The mirror neuron system

Christian Keysers, Marc Thioux, and Valeria Gazzola (Chapter 14) provide a review of the mirror neuron system (MNS) in social cognition, in both monkeys and its putative equivalent in humans. They argue for this being a building block of major human abilities, from imitation to language. Giacomo Rizzolatti and Maddalena Fabri-Destro (Chapter 15) provide their own first-hand perspective on the discovery of the MNS in the monkey brain, and their view of how the human MNS is dysfunctional in autism.

Oxytocin

Markus Heinrichs, Frances Chen, and Gregor Domes (Chapter 16) report on the latest research into the role of the peptide hormone oxytocin (OXT) in our capacity for empathy and social cognition. They argue that OXT increases social approach by reducing social stress reactivity (evidence of lower cortisol during OXT administration), and boosting social motivation. They argue that the amygdala is the target for OXT. Amygdala volume and activation during emotion processing are also correlated with polymorphisms in the oxytocin receptor gene (*OXTR*) and OXT dampens amygdala stress responses. OXT also increases interpersonal trust as well as attachment, and *OXTR* variations are associated with maternal sensitivity to their child's needs. OXT also boosts performance on ToM/emotional accuracy tests, and increase amount of gaze to the eye region of the face. This paints an important picture of how OXT sets the stage for focusing on another's mental states, and for learning to use a ToM. The authors explore the potential of OXT for therapy for conditions involving social anxiety and deficits in social cognition.

Prenatal testosterone

Bonnie Auyeung and Simon Baron-Cohen (Chapter 17) review work indicating that fetal testosterone (FT) is inversely associated with a range of indicators of social development, such as eye contact, language development, mentalizing, and empathy. They also report on a recent study by Mike Lombardo and colleagues showing that FT is associated with increased gray matter volume of the RTPJ, a key mentalizing brain region. They argue these effects are specific to the prenatal effects of testosterone, and report positive associations between FT levels and the number of autistic traits found later in development. However, they also review evidence that administration of testosterone in adulthood changes activation levels of a number of brain regions relevant to ToM and emotion processing, such as orbitofrontal cortex (OFC) and amygdala, as well as reward circuitry, such as the ventral striatum. They argue that testosterone may have “opposite” effects to oxytocin.

Genetics

Bhismadev Chakrabarti and Simon Baron-Cohen (Chapter 18) discuss the heritability of empathy using evidence from twin studies. They also discuss different approaches to identifying the genetic basis of autism, in which ToM is impaired. These approaches include genome-wide association studies, copy number variations, and candidate single nucleotide polymorphisms (SNPs). They adopt the latter approach by studying SNPs in genes involved in neural growth and connectivity, or in social and emotional responsivity, and in sex steroid hormones. Genes associated with empathy included *NTRK1* and *NTRK3*, *ESR2*, *GABRB3*, and *OXTR* among others. They make a case for taking a systems-based approach to understanding the function of genes that might relate to empathy and ToM.

Deaf children

Jennie Pyers and Peter de Villiers (Chapter 19) summarize the development of ToM in deaf children raised by signing parents (so-called deaf children born to deaf parents (DoD)) vs. deaf children born to hearing parents (DoH) and who are orally taught, to tease out the role of language in the development of ToM. They report how deafness per se does not impact ToM development since deaf children brought up as native signers perform as well as typically

hearing children. However, deaf children brought up by hearing parents show language delay and subsequent delay in the development of ToM. This clearly illustrates the role that language plays in ToM development. Other studies reveal the complex interplay between language and ToM in the deaf, and connect with Wellman and Peterson's earlier chapter dealing with this question. They explore the important question about the role of establishing joint attention in children who are deaf, and whether this is a critical mediating factor in whether ToM proceeds typically or not.

Psychopaths

James Blair and Stuart White (Chapter 20) remind us that, of all clinical groups, those with anti-social personality disorder, a subset of whom would meet criteria for psychopathy, are the clearest case of a group who lack emotional empathy, despite having excellent cognitive empathy and ToM. They can manipulate and even torture a victim by knowing very well what their victims thoughts and feelings are, but don't have the typical emotional responses to another person's suffering. They describe their "integrated emotions systems" model of how a typical child learns morality, the key role of the amygdala, insula, and inferior frontal cortex (IFC) in this process, because these brain regions are critical for forming associations with negative emotions, such as fear, disgust, and anger; and the role of the vMPFC in moral decision-making. The pattern of empathy deficits in psychopaths makes them a kind of mirror-image of those with autism, who struggle with cognitive empathy but may have intact emotional empathy.

Autism

Antonia Hamilton and Lauren Marsh (Chapter 21) devote their chapter to ToM in autism, hinted at frequently in other chapters in this volume, but central to this one. They focus on the mirror system in typical ToM, particularly the IFG and the anterior intraparietal sulcus (aIPS) in decoding others' actions. They contrast this with the brain's mentalizing system, particularly TPJ and mPFC. They explore the evidence for each of these two theories: the "broken mirror" theory vs. the impaired mentalizing theory of autism. Although early work found evidence supportive of the "broken mirror" theory, subsequent studies have found no differences during observation and imitation of other's actions. Studies from Hamilton's laboratory contrast atypical mentalizing system activity in autism, particularly in the mPFC, to intact mirror system engagement.

Peter and Jessica Hobson (Chapter 22) tackle the slippery concept of "self" in autism, reviewing studies of self-awareness, self-conscious emotions (particularly guilt and embarrassment), and reflection on one's own mental states, use of the first-person pronoun, and the self-reference effect in memory, all of which point to difficulties in the development of a concept of self and the self-monitoring function in autism. They review evidence from fMRI studies consistent with the view that in autism the self is atypical. Peter Carruthers (Chapter 26) provides a philosopher's perspective on self- vs. other-directed use of ToM. Julie Hadwin and Hanna Kovshoff (Chapter 23) usefully review teaching methods and interventions targeting ToM deficits in autism. These methods range from didactic approaches to breaking down ToM into principles, through to facilitating joint attention as a precursor to ToM, through to autism-friendly methods of teaching emotion recognition. These chapters are important in linking the nature of autism to clinical and educational practice.

Non-human primates

Andrew Whiten (Chapter 24) reminds us that whilst humans are “inveterate mentalists”, our “baroque human mental interpenetration is unparalleled in its complexity and depth”. His chapter demonstrates that non-human primates have some elementary aspects of ToM, and argues that to understand this remarkable human achievement, we need an evolutionary framework. He reminds us that agriculture is only 10 000 years old, and that the evolutionary landscape to which we adapted was a hunter-gatherer lifestyle with a home base. Whiten reminds us that this niche was uniquely human—no other ape developed it. He retells the standard story of how, following the loss of forest cover in Africa, humans had to adapt by becoming bipedal and venturing out into the open savannah, having to outwit dangerous predators and become big-game hunters. Apes, in contrast, stayed in the forest. Humans alone had to develop the intelligence of using weapons and traps (requiring deception and ToM) instead of teeth and claws.

Whiten disputes the standard story as overlooking key factors in human evolution from studies of modern-day hunter-gatherers. This hints that our human ancestors probably lived in communities that were egalitarian and cooperative; how the base-camp likely involved information-sharing and a division of labour between the sexes; how hunting is akin to being a group-predator, rather than an individual predator; how hunter-gatherers developed culture and language; and how ToM fits into this “socio-cognitive niche”. He also reviews the primate ToM literature over 30 years since Premack and Woodruff asked if the chimpanzee has a ToM, concluding (with Tomasello and Call) that chimpanzees may understand goals, intentions, perceptions, and even the knowledge states of others, but that they do not understand other’s beliefs. This was the Rubicon that humans alone crossed.

These 26 chapters represent, for us as editors, a wonderful overview of a field that is as exciting today as it was when we published *UOM-1* in 1993. We thank our contributors and look forward to meeting them as authors and you as readers again, in *UOM-4*!

Acknowledgements

We are grateful to our contributors for their hard work in producing these outstanding chapters, and for working patiently with us as editors and with the staff at Oxford University Press over a 2-year period, through a big production process that has resulted in such a high quality volume.

We dedicate this volume to Professor Donald Cohen MD (September 5, 1940–October 2, 2001) who was an enthusiastic co-editor of the first two editions of this book and director of the Yale Child Study Center and the Sterling Professor of Child Psychiatry, Pediatrics, and Psychology at the Yale School of Medicine. He made fundamental contributions to the understanding of autism and Tourette's syndrome, and was a passionate advocate for social policy. His multi-disciplinary interests crossed psychology and biology, and included the clinical importance of "theory of mind".

This page intentionally left blank

Contents

Contributors xv

Section 1 **Development and cognition**

- 1** Early manifestations of mindreading 3
Victoria Southgate
- 2** Learning about the mind from evidence: Children's development of intuitive theories of perception and personality 19
Andrew N. Meltzoff and Alison Gopnik
- 3** Teleology: Belief as perspective 35
Johannes Roessler and Josef Perner
- 4** Theory of mind, development, and deafness 51
Henry M. Wellman and Candida C. Peterson
- 5** Can theory of mind grow up? Mindreading in adults, and its implications for the development and neuroscience of mindreading 72
Ian Apperly
- 6** Mind attribution is for morality 93
Liane Young and Adam Waytz
- 7** Issues in the measurement of judgmental accuracy 104
David A. Kenny

Section 2 **Neural systems and mechanisms**

- 8** Brain electrophysiological studies of theory of mind 119
Mark A. Sabbagh
- 9** Functional neuroimaging of theory of mind 132
Jorie Koster-Hale and Rebecca Saxe
- 10** Theory of mind: Insights from patients with acquired brain damage 164
Dana Samson and Caroline Michel
- 11** Understanding emotional and cognitive empathy: A neuropsychological perspective 178
Anat Perry and Simone Shamay-Tsoory
- 12** Empathy and the brain 195
Cade McCall and Tania Singer
- 13** Neural sources of empathy: An evolving story 214
Jamil Zaki and Kevin Ochsner

- 14 Mirror neuron system and social cognition 233
Christian Keysers, Marc Thioux, and Valeria Gazzola
- 15 The mirror mechanism: Understanding others from the inside 264
Giacomo Rizzolatti and Maddalena Fabbri-Destro
- 16 Social neuropeptides in the human brain: Oxytocin and social behavior 291
Markus Heinrichs, Frances S. Chen, and Gregor Domes
- 17 Prenatal and postnatal testosterone effects on human social and emotional behavior 308
Bonnie Auyeung and Simon Baron-Cohen
- 18 Understanding the genetics of empathy and the autistic spectrum 326
Bhismadev Chakrabarti and Simon Baron-Cohen

Section 3 **Psychiatric, neurodevelopmental, and neurological disorders**

- 19 Theory of mind in deaf children: Illuminating the relative roles of language and executive functioning in the development of social cognition 345
Jennie Pyers and Peter A. de Villiers
- 20 Social cognition in individuals with psychopathic tendencies 364
James Blair and Stuart F. White
- 21 Two systems for action comprehension in autism: Mirroring and mentalizing 380
Antonia Hamilton and Lauren Marsh
- 22 Autism: Self and others 397
Peter R. Hobson and Jessica A. Hobson
- 23 A review of theory of mind interventions for children and adolescents with autism spectrum conditions 413
Julie A. Hadwin and Hanna Kovshoff

Section 4 **Comparative and philosophical perspectives**

- 24 Culture and the evolution of interconnected minds 431
Andrew Whiten
- 25 Mindreading by simulation: The roles of imagination and mirroring 448
Alvin I. Goldman and Lucy C. Jordan
- 26 Mindreading the self 467
Peter Carruthers

Index 487

Contributors

Professor Ian Apperly

School of Psychology, University of Birmingham, UK

Dr Bonnie Auyeung

Autism Research Centre, Department of Psychiatry, University of Cambridge, UK

Professor Simon Baron-Cohen

Autism Research Centre, Department of Psychiatry, University of Cambridge, UK

Dr James Blair

Unit on Affective Cognitive Neuroscience at NIMH, USA

Professor Peter Carruthers

Department of Philosophy, University of Maryland, USA

Dr Bhismadev Chakrabarti

School of Psychology and Clinical Language Sciences, University of Reading, UK

Dr Frances S. Chen

Department of Psychology, University of Freiburg, Germany

Professor Peter de Villiers

Department of Psychology, Smith College, USA

Dr Gregor Domes

Department of Psychology, University of Freiburg, Germany

Dr Maddalena Fabbri-Destro

Brain Center for Motor and Social Cognition, Italian Institute of Technology, Department of Neuroscience, Italy

Dr Valeria Gazzola

Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands

Professor Alvin I. Goldman

Department of Philosophy, Center for Cognitive Science, Rutgers, State University of New Jersey, USA

Professor Alison Gopnik

Department of Psychology, University of California at Berkeley, USA

Dr Julie A. Hadwin

School of Psychology, University of Southampton, UK

Dr Antonia Hamilton

School of Psychology, University of Nottingham, UK

Professor Markus Heinrichs

Department of Psychology, University of Freiburg, Germany

Dr Jessica A. Hobson

University College London, Institute of Child Health (ICH), UK

Professor Peter R. Hobson

University College London, Institute of Child Health (ICH), UK

Lucy C. Jordan

Department of Philosophy; Rutgers, State University of New Jersey; New Brunswick, USA

Professor David A. Kenny

Department of Psychology, University of Connecticut, USA

Professor Dr Christian Keyzers

Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands

Jorie Koster-Hale

Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, USA

Dr Hanna Kovshoff

School of Psychology, University of
Southampton, UK

Lauren Marsh

School of Psychology, University of
Nottingham, UK

Dr Cade McCall

Department of Social Neuroscience,
Max Planck Institute for Human Cognitive
and Brain Sciences, Leipzig, Germany

Professor Andrew N. Meltzoff

Institute for Learning and Brain Sciences,
University of Washington, USA

Dr Caroline Michel

Psychological Sciences Research Institute
Université catholique de Louvain, Belgium

Professor Kevin Ochsner

Department of Psychology, Columbia
University, USA

Dr Josef Perner

Department of Psychology and
Centre for Neurocognitive Research,
University of Salzburg, Austria

Dr. Anat Perry

Department of Psychology, University of
Haifa, Israel

Professor Candida C. Peterson

School of Psychology, University of
Queensland, Australia

Dr Jenny Pyers

Department of Psychology,
Wellesley College, USA

Professor Giacomo Rizzolatti

Department of Neuroscience,
University of Parma, Italy Parma, Italy

Dr Johannes Roessler

Department of Philosophy,
University of Warwick, UK

Professor Mark A. Sabbagh

Psychology Department, Queen's University
at Kingston, Canada

Professor Dana Samson

Psychological Sciences Research Institute
Université catholique de Louvain, Belgium

Dr Rebecca Saxe

McGovern Institute for Brain Research and
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, USA

Dr Simone Shamay-Tsoory

Department of Psychology, University of
Haifa, Israel

Professor Tania Singer

Department of Social Neuroscience,
Max-Planck-Institute, Institute for
Human Cognitive and Brain Sciences,
Leipzig, Germany

Dr Victoria Southgate

Centre for Brain and Cognitive Development,
School of Psychology, Birkbeck,
University of London, UK

Dr Marc Thioux

Netherlands Institute for Neuroscience,
Royal Netherlands Academy of Arts and
Sciences, Amsterdam, The Netherlands

Dr Adam Waytz

Management and Organisations Department,
Kellogg School of Management, Northwestern
University, USA

Professor Henry M. Wellman

Department of Psychology, University of
Michigan, USA

Dr Stuart F. White

Unit of Affective Cognitive Neuroscience
National Institute of Mental Health/NIH, USA

Professor Andrew Whiten

School of Psychology and Neuroscience,
University of St Andrews, UK

Dr Liane Young

Department of Psychology,
Boston College, USA

Dr Jamil Zaki

Department of Psychology, Stanford
University, USA

This page intentionally left blank

Section 1

Development and cognition

This page intentionally left blank

Chapter 1

Early manifestations of mindreading

Victoria Southgate

A decade ago, there was no chapter within a book on understanding other minds devoted to the infancy period, because it was generally agreed that no such understanding existed at this stage of development. Traditional tests of mindreading, like the Sally-Anne or Smarties task (Gopnik & Astington, 1988; Baron-Cohen, Leslie & Frith, 1985; Perner, Leekam & Wimmer, 1987; Wimmer & Perner, 1983), were robustly failed by children under the age of 4, and it was concluded that, until this point, children essentially lacked an appreciation that other's behavior can be driven by unobservable mental states (Wellman, Cross & Watson, 2001). While some authors maintained that younger children's failure on these classic tasks likely reflected performance issues, rather than any conceptual deficit (Leslie, 1994, Leslie & Polizzi, 1998), modifications aimed at enabling young children to overcome performance limitations did not lower the age at which children passed these tasks to any great degree (e.g. German & Leslie, 2000). Nevertheless, it was puzzling that, despite failing these tasks, young children's behavior suggested a greater understanding of other minds than their explicit task performance would give them credit for. For example, one influential theory of communication holds that it involves the ability to represent others' mental states (Grice, 1989). This is clear when we consider the case of pronouns: using terms like "it", or "he" suggests that the person using those terms considers that the other person shares their understanding of what "it" is, or who "he" is, and these pronouns are abundant in the speech of toddlers (Bloom, 2000).

It has long been known that the method used to test for the presence of an understanding has an influence on whether or not such understanding is revealed. For example, relying on infants explicit search behavior would lead us to believe that they have no appreciation of object permanence before 7 months, when they begin to uncover hidden objects themselves (Piaget, 1954). However, researchers using looking-time (the length of time that infants look toward outcomes that are expected or unexpected if one possesses a particular concept), and anticipatory looking have demonstrated that, in fact, infants as young as 3 months grasp object permanence (Baillargeon, Spelke & Wasserman, 1985; Ruffman, Slade & Redman, 2005). In a move paralleling that previously seen in the domain of physical cognition, the last decade has seen researchers turn to infants looking behavior, as a measure of their social cognitive abilities. In findings paralleling those in the domain of physical cognition, infants looking behavior reveals a far greater appreciation that other's behavior is generated by their own representations of the world, than classic tests of mindreading would have us believe.

Evidence for mindreading in infancy

The first forays in to using infants looking behavior as an indication of what they understand about other minds investigated whether infants understand that others' movements are directed toward their goals (e.g. Gergely, Nadasdy, Csibra, & Biro, 1995; Woodward, 1998). In one of the most well replicated results in the domain of infant social cognition, Woodward (1998, 1999)

demonstrated that 6- and 9-month-old infants encoded a relationship between an agent and the target of its actions, and reacted with longer looking when the agent subsequently acted on a new target object. While these findings demonstrate that infants encode actions in terms of what they are directed toward (their intentionality; Gomez, 2008), and expect that having acted on a target multiple times, people will most likely continue to act on the same object again, they do not tell us whether infants are appealing to unobservable mental states to generate these expectations. There is no need to attribute to infants an understanding that others' actions are generated by unobservable mental states like "goal", as alternative non-mentalistic interpretations are equally plausible (Gergely & Csibra, 2003; Ruffman, Taumoepeau & Perkins, 2011). Thus, evidence that young children are using others' mental states in order to generate expectations about their behavior needs to come from situations in which the child and the other hold a different representation of the world (Dennett, 1978), such as is the case in the classic false-belief task.

False belief understanding in infants

In the first study to suggest that infants are capable of representing other's unobservable mental states, Onishi & Baillargeon (2005) cleverly transformed the classic Sally-Anne task (Baron-Cohen et al., 1985; Wimmer & Perner, 1983) in to a looking-time paradigm. Infants observed as an actor placed an object in to one of two boxes and then reached in to retrieve her object. After that, infants saw the agent disappear, during which time the object moved, by itself, to the opposite box. Infants then observed one of two outcomes. In the congruent test trial, the actor then reached in to the box where she had left her object (congruent because she could not have known that the object had moved in her absence and so she should reach in to the box where she left it), but in the incongruent test trial, she reached in to the box where the object actually is (incongruent because, unaware that the object had moved, she should not reach to its real location). Onishi & Baillargeon (2005) found that infants look significantly longer when they watch the incongruent outcome than the congruent outcome, suggesting that they find someone searching in the location where the object really is (when they could not know where it really is) unexpected—and the fact that they respond with longer looking times toward an outcome in which the person searches where the object actually is, suggests that infants have some understanding that people's actions should depend on what they have experienced, rather than on what is actually the case.

One of the objections to the conclusions drawn by Onishi & Baillargeon (2005), that their finding demonstrates that infants understand that others act on the basis of their representations, was that any understanding of mental states should manifest in a variety of contexts, and that we are only permitted to conclude that 4-year-olds operate with a mentalistic understanding of others because such understanding has been demonstrated in many different contexts and scenarios (Perner & Ruffman, 2005). In answer to this challenge, many studies have now appeared which do extend this original finding to different contexts. For example, Song and Baillargeon (2008) showed that infants understand that others can have representations concerning not only the location of an object, but also its contents. In a non-verbal task reminiscent of the classic "Smarties" task (Perner et al., 1987), they showed infants that an actor preferred to reach for a doll with blue hair than a skunk. Then, when the actor was absent, someone put the doll in a plain box and the skunk in a box that had blue hair sticking out of it. Like children who pass the Smarties test, 14.5-month-old infants seemed to expect the actor to be misled by the visible tuft of hair and looked longer when she searched in the plain box, even though that was actually where her desired

doll was located. There are now numerous other examples of different contexts in which infants appear able to think about what other people have experienced (e.g. Luo, 2011; Song, Onishi, Baillargeon & Fisher, 2008; Scott, Baillargeon, Song & Leslie, 2010; Surian, Caldi, & Sperber, 2007; Yott & Poulin-Dubois, 2012).

What kind of behavior do these representations support?

Looking-time studies have demonstrated that infants are sensitive to the fact that others' actions are motivated by their own view of the world, even when this is different from what the infant should know to be the real state of the world. In other words, preverbal infants appear to represent another person's representation of the world. What kind of behavior might this ability to represent other's perspectives on the world support? It is often held that one of the primary reasons why we may need to engage in mindreading is to enable us to generate predictions concerning what other people may do (Premack & Woodruff, 1978; Dennett, 1978), and the ability to accurately predict others' actions is crucial for any social species (Verfaillie & Daems, 2002). However, because measures like looking-time rely on infant's responses to the outcomes of events, they cannot tell us whether infants are able to use this sensitivity to others' mental states, to generate accurate predictions about their behavior. While looking time is assumed to rest on violated expectations (or predictions) that are formed in advance of the infant seeing an outcome, it may equally reflect the infant's recognition that an outcome is incongruent with a preceding event only when the outcome is seen (Southgate & Csibra, 2009). Thus, while looking time certainly tells us that infants can recognize behavior that is inconsistent with a perspective that an agent should hold, it does not tell us whether infants can use their sensitivity to others' perspectives usefully, to generate predictions about what they may consequently do.

In a first step toward addressing this question, we used eye-tracking to ask whether 25-month-old infants could predict where an agent will search for a desired object when she has a false belief about that object's location (Southgate, Senju & Csibra, 2007). In this task, infants first saw two familiarization trials in which an agent, sitting behind a panel with two windows and two boxes (Figure 1.1), observed as a puppet placed an object in to one of the two boxes. The puppet then disappeared, and both windows lit up at the same time as a chime was heard—at which point the agent reached through the window located behind the box containing the toy, and opened the box. These two familiarization trials served to show the infant that this cue (both windows lighting up and chime) signaled the imminent reach of the agent through one of the windows. In the test trial, infants then saw a false-belief scenario in which, once the puppet had placed the toy in one of the boxes, the agent turned around, and did not witness the puppet subsequently remove the toy from the box and take it away. The question was, when the agent turned back toward the boxes, could infants use their apparent sensitivity to others' perspectives on events to predict that the agent will reach through the window above the box where she saw the toy being placed, even though the infant has seen that the toy is no longer there? Results showed that, indeed, they could. When the agent turned around after the toy was placed in the left-hand box, infants predicted she would open the left-hand window upon her return, and when she turned around after it was placed in the right-hand box, they expected her to open the right-hand window.

More recently, Knudsen & Liszkowski (2012a) have also demonstrated that even younger infants, at 18 months, generate predictions in accord with others' epistemic states. In their study, they use what they call an "anticipatory intervening" paradigm, originally suggested by Dennett (1978), in which infants observed as an actor re-entered a room with either a true- or false-belief about the location of her toy. When the actor had a false-belief that her toy was in the box where she

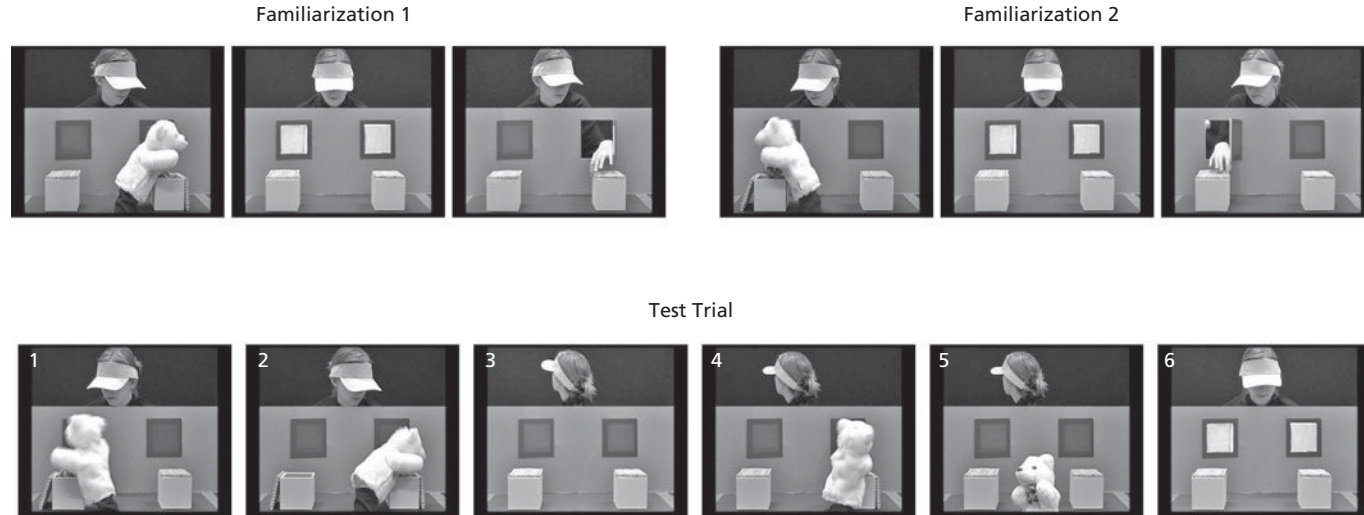


Figure 1.1 Upper panel: The two familiarization trials in which the puppet places a ball in either the right or left box. After the puppet leaves the scene, the two windows are illuminated and shortly after the actor reaches through the window above the box in which the puppet has left the ball. Lower panel: Test trial. The puppet places the ball in the left-hand box (1) and then removes it and places it in the right-hand box (2). The actor then turns around (3) and the puppet returns to remove the ball from the right-hand box (4 and 5). The actor then reorients to the scene and the windows are illuminated. Infants looking behaviour is measured from the onset of the illumination. A second group of infants received an alternative test trial in which the correct location for the actor to search was the left-hand box.

left it, 18-month-olds intervened to warn the actor of the real location of the toy, before she had approached the location where she thought the toy was, but they did not do this when she had a true belief. Together, these studies suggest that, by 18 months, infants can use their understanding that others' actions are modulated by their epistemic states, to generate predictions about what actions will follow, and even prepare an appropriate response (e.g. pointing to inform the other) for that prediction.

Mindreading or behavior reading?

The data reviewed so far have been interpreted as reflecting infant's appreciation that unobservable mental states modulate others' actions (e.g. Baillargeon, Scott, & He, 2010). However, a constant challenge for those advocating that infants are encoding others' mental states, is that simpler, non-mentalistic interpretations in terms of behavior-reading are always possible (Low & Wang, 2011; Ruffman et al., 2011). The problem is that one cannot know for sure whether infants are genuinely using mental state concepts such as "knows" or "believes" when generating predictions about what the other will do, or whether they have simply learnt associations between events and typical behaviors that follow from those events (Ruffman et al., 2011). Since mental states drive behavior, every mental state will have a behavioral correlate. One commonly suggested probability that children might exploit is that people who are ignorant of something will usually get the wrong answer (e.g. Perner & Ruffman, 2005). If children evoke such an "ignorance rule" in the false-belief task (e.g. the Sally-Anne task), they would reason that, because Sally did not see Anne move the toy from location A to location B, Sally will probably get the wrong answer (i.e. she will choose location A because the infant knows that the object is actually in location B). However, crucially, this does not entail that the child attributes to Sally a mental representation that the toy is at location A—she will simply search there because this is not where the object is. While the evidence that children evoke such a rule is somewhat contradictory (e.g. Friedman & Petrashek, 2009; Hedger & Fabricius, 2011; Ruffman, 1996), it does not seem that infants are using this rule in solving these tasks. In our previously described study (Southgate et al. 2007), infants saw a scenario in which the object was actually removed from the scene altogether by the puppet (rather than being transferred from A to B), rendering both boxes the wrong place to search if you actually want to find the ball. Infants do not expect the agent to search indiscriminately, however; they specifically expect her to search in the location where she last saw the object, suggesting that they attribute to her an epistemic state in which the object is in location A. Similarly, in a follow-up to their original study (Knudsen & Liszkowski, 2012a), Knudsen & Liszkowski (2012b) again used their "anticipatory intervening" paradigm to tease apart responses based on an ignorance rule and those based on an attributed epistemic state. As in Southgate et al., (2007), the agent's toy was removed entirely during the period when the agent was absent and, instead, aversive objects that the infant knew the agent did not like, were placed in each container. When the agent returned, infants pointed more to the location where the agent had left her toy, suggesting that they expected her to specifically choose one location. If infants were operating with an ignorance rule, since both locations were wrong (as the object had been removed, rather than transferred), they should have pointed to both locations. In a final condition, the agent was, in fact, ignorant of the location of her toy and, in this case, infants pointed equally to both locations, suggesting that they understood that her ignorance should lead her to search indiscriminately. Recently, looking-time studies have also begun to include conditions that control for the use of such an ignorance rule, and confirm that infants have different expectations depending on whether the agent should have a specific (false) belief about the location of the object, or is simply ignorant of its location (e.g. Scott & Baillargeon, 2009).

While infants do not seem to be relying on an ignorance rule, nevertheless, there are other behavioral rules or probabilities that infants might have acquired. For example, in making their case against mindreading in non-human primates, Povinelli & Vonk (2003) argue that, while these animals might seem like they are sensitive to others' mental states, they may simply have learned that people tend to search for things in locations toward which they were last oriented (see Perner & Ruffman, 2005 for the suggestion that infants may exploit the same rule). For Povinelli & Vonk (2003), learning that orientation is correlated with subsequent search is sufficient to form the expectation that people will search for things where they last saw them. So, if an actor is oriented toward the left-hand box when the object was placed in it, and this was the last point at which her eyes were oriented toward the object, this is where she will probably search for it. Since we usually do look for things where we last put them, or saw them, it is plausible that infants might acquire such a rule through observing the behavior of others.

Of course, this kind of rule- or probability-based explanation applies equally to the standard false-belief task which is taken as evidence that older children operate with a representational theory of mind, and are not just good behavior readers. However, as older children can, in addition to generating correct predictions, also explain why Sally will search in the false-belief location, and give verbal explanations consisting of mental state terms like "thinks" and "knows," we can be more confident that they are not solving the tasks merely by employing these kind of behavioral rules. How could we know whether infants are really considering the unobservable mental states of others in generating predictions about their behavior, or whether these predictions are based on associations between events and behavior?

One strategy has been to argue that, because infants succeed on a whole host of different tasks, and each would require a different learned association to be solved by applying different behavioral rules, it is simply more parsimonious to assume that they are solving these tasks by appealing to common mental states (Baillargeon et al., 2010).¹ Another strategy has been to directly test whether young children inflexibly use such rules. For example, Trauble, Marinovic, & Pauen (2010) directly tested the hypothesis that infants solve non-verbal false-belief tasks by using the "people search in the last place that they saw something" rule. To address this, Trauble and colleagues (2010) ran a study that was largely similar to Onishi and Baillargeon (2005), with the exception that, in one condition (the manual control condition) the agent gained their knowledge through touch, rather than vision. While the agent was turned around, she manipulated a balance beam behind her back, which caused the object to roll in to the opposite box. Thus, even though she did not actually see the object move (and so the last place she saw the object was not the location where it now is), her own manipulation caused the object to move, and so she should know its actual location. As Trauble and colleagues (2010) argue, if infants rely blindly on a rule like "people search for things where they last saw them," the agent should still be expected to look for her object in the location where she last saw it, even though she should know that it is in the opposite box since it was her who caused its movement. Infants are not fooled by this: they do not expect her to search for the ball in the location where she last saw it, instead they look longer when she searches in this location than in the location where the object actually is (just as infants in a true-belief condition did), presumably because they understand that the agent should know where the ball actually is.

Nevertheless, Povinelli & Vonk (2003) are right when they argue that "no experiment in which theory of mind coding derives from a behavioral abstraction will ever suffice" to definitely answer

¹ See Perner (2010) for arguments against this position.

the question of whether non-verbal subjects are utilizing behavioral rules, or appealing to other's mental states. Even if infants in Trauble and his colleagues' (2010) study were not relying on a "seeing" rule, they may have been relying on some other rule involving "touching" (Low & Wang, 2011). Povinelli and Vonk's point is that, because every mental state will have a behavioral correlate, any study in which mindreading skills are assessed by comparing performance on conditions which vary in a behavior (e.g. true belief vs. false belief), can always be explained in terms of rules. Since we simply do not know how much experience infants need in order to form behavioral correlations, appealing to parsimony to solve this dilemma is unsatisfying. However, Povinelli & Vonk (2003) do advocate a possible solution following an earlier suggestion by Heyes (1998). Specifically, in order to tease apart genuine appreciation of unobservable mental states from reliance on behavioral associations, one needs to create a situation in which expectations are based on an experience for which it would not have been possible for the subject to have acquired any behavioral correlate. This could be achieved by providing the subject with direct first-person experience of one of its own mental states—seeing—and asking whether they can extrapolate this subjective experience to another individual, even though they would have no opportunity to acquire any behavioral correlate of this mental state. A version of this paradigm was recently implemented by Meltzoff and Brookes (2008) in which they provided 12- and 18-month-old infants with differential experience with a blindfold. For one group of infants, the blindfold was opaque and prevented seeing; for the other group, a trick blindfold which, although perceptually identical to the opaque blindfold, did allow the infant to see. They tested infants abilities to extrapolate their first-person experience with the blindfold to others, by asking whether, when infants saw another person wearing the blindfold (all infants actually saw the experimenter wear the opaque blindfold), they would selectively follow the "gaze" direction of that person, depending on which blindfold they had experienced. Their results showed that infants who had experienced the trick blindfold subsequently followed the gaze of the experimenter who was wearing the blindfold significantly more than infants who had experienced the opaque blindfold, suggesting that, indeed, infants had extrapolated their own first-person experience of seeing, or not seeing, to another person.

In a recent study, we extended this methodology to a false-belief scenario with 18-month-olds, and asked whether differential blindfold experience not only modulated what infants understood about what the other could see, but also their predictions about what the other would do (Senju, Southgate, Snape, Leonard & Csibra, 2011). As in Meltzoff and Brooks (2008), we provided two groups of infants with either an opaque blindfold experience, or a trick (transparent) blindfold experience. Following the blindfold experience, all infants took part in an anticipatory looking paradigm similar to that used by Southgate et al. (2007). The main difference between this study and the previous study is that, while in Southgate et al. (2007) infants saw the agent turn around while the puppet removed the object from the box, in this version the agent donned the blindfold.² We hypothesized that if infants understood from their own experience that the blindfold was opaque or transparent, then this should modulate whether they infer that the agent has seen the puppet remove the object or not and, consequently, whether the agent will search in the location where the object was before she put on the blindfold. Specifically, we hypothesized that infants who had the opaque blindfold experience would understand that the agent could not see the puppet removing the object, and so should expect her to search in the location where she last

² Both the transparent and opaque blindfold looked identical when worn by someone else, but regardless of condition, all infants saw a movie in which the agent wore the opaque blindfold.

saw the object (the false-belief location), whereas we hypothesized that infants who had the trick blindfold experience would understand that the agent could see the puppet removing the ball and so would have no particular expectation about where she would search. This is exactly what we found. Fourteen out of 18 infants who had undergone the opaque blindfold experience expected the agent to search in the false-belief location (a result significantly above chance), whereas only six out of 18 infants who had the trick blindfold experience looked in anticipation toward the false-belief location. These results show that, not only did infants use self-experience to assess what the other person could see, they understood the causal role that seeing has in generating action. Crucially, these results cannot be explained in terms of a reliance on behavioral rules because, since infants never saw the adult wearing the blindfold, they had no opportunity to acquire any behavioral correlate of blindfold wearing.³

Beyond behavioral rules

The evidence reviewed above suggests that infants are not solving these false-belief scenarios by a reliance on learned behavioral rules and that they are, indeed, representing the world from the perspective of the other. However, to what extent does infants' success on these tasks reflect an appreciation of mental state concepts as they are represented by adults, and what might infants' representations look like? One of the biggest challenges is to account for why, if infants do possess a genuine representational theory of mind, do children younger than 4 years not pass the standard false-belief tasks?

Several theorists have attempted to explain this paradox. One theory is that young infants have a representational theory of mind, with concepts like seeing, knowing and believing, but are unable to demonstrate this knowledge on the standard false-belief tasks (German & Leslie, 2000; Baillargeon et al., 2010). For example, Baillargeon and colleagues argue that in order to respond correctly on the verbal false-belief task, infants would need not only to represent the other's false belief, but additionally they would need to access this representation in order to select the correct response, and simultaneously inhibit a prepotent tendency to answer the question based on what they know to be true (Baillargeon et al., 2010; Scott & Baillargeon, 2009). This position advocates that infant's representations of other minds are not different from adults, but due to limited resources, or immature neural connections, they cannot do all three things (represent beliefs, select the right answer and inhibit the wrong answer) at once.

There are several problems with this account. Fundamentally, this account is open to similar criticisms as previous accounts that involved explaining infants' failures by appealing to limited executive resources. By designing tasks that alleviate some of the executive demands, several researchers have succeeded in lowering the age of passing false-belief tasks by a few months, but there is no dramatic change (e.g. Carlson, Moses & Hix, 1998; Surian & Leslie, 1999). A bigger problem is that it is not clear why performance on the so-called "spontaneous-response tasks" should not also require these additional elements that Baillargeon and colleagues (2010) reserve for the elicited response tasks. For Baillargeon and colleagues (2010), infant's performance on their non-verbal false-belief tasks suggest that infants "realize that others act on the basis of their beliefs and that these beliefs are representations that may or may not mirror reality" (Onishi & Baillargeon, 2005). If it is, indeed, the case that infants' performance on non-verbal false-belief tasks reflects this kind of understanding

³ See Penn & Povinelli (2007) for further arguments as to why expectations modulated by self-experience necessarily entails mental state attributions.

(including an appreciation that the agent's belief is a representation that is false), then presumably some response inhibition would need to be deployed in order for the infant not to respond based on what the infant knows to be the true location of the object? In Baillargeon and colleagues' (2010) response account, it is not clear why a saccadic response (e.g. Senju et al., 2011; Southgate et al., 2007) should be immune to an egocentric bias, but a verbal response should not be.

A second problem for this account is that more recent paradigms, in which infants also appear to consider others false beliefs, would appear to involve the elements of response selection and response inhibition that Baillargeon and colleagues (2010) argue are too challenging for infants' limited resources. In one task, Southgate, Chevallier and Csibra (2010) presented 17-month-old infants with a paradigm in which, in order to select the object about which the experimenter was referring, infants needed to consider where the experimenter believed the object to be. Infants saw the experimenter placing a different novel object, one in each of two similar boxes, and closed the lids. The experimenter then left the room, after which a new experimenter suddenly appeared from behind the curtains at the back of the room, crept over to the boxes and switched their contents (i.e. the object in the left-hand box was transferred to the right-hand box, and vice versa). The new experimenter then closed the lids again and crept back behind the curtains. At this point, the original experimenter returned to the room, knelt down behind the two boxes and pointed toward one of the closed boxes and said to the infant "Do you remember what I put in here? Can you get it for me?"⁴ Since the experimenter was not in the room when the objects were switched, infants should reason that she is *not* now referring to the object that is actually in the box toward which she is pointing, but the object that was in the box and is now in the other box. So, the correct response is to give the experimenter the object in the non-referred-to box. The results showed that this is what infants did: nine out of 12 infants selected the object in the non-referred box in response to the experimenter's request. On the other hand, in a true-belief control condition in which the first experimenter stayed in the room and witnessed the second experimenter switching the contents of the boxes, nine out of 12 infants chose the box that the experimenter was actually pointing to. In this case, because she saw the switch and knew the contents of the boxes, infants took her pointing as referring to the object that was now in the box.

According to Baillargeon and colleagues, this kind of "indirect elicited-response" task would not demand the same kind of response selection and inhibition processes required by the standard false-belief task because there is no question that directly taps their representation of the agent's belief (Baillargeon et al., 2010). It is not clear on what grounds Baillargeon and colleagues (2010) draw the conclusion that a direct (e.g. standard Sally-Anne task) and an indirect (e.g. Southgate et al., 2010) elicited-response task would require differences in response-selection and response-inhibition. Both tasks require an explicit⁵ response from the infant, and in order to generate their response, infants need to consider what the agent thinks about the location of the objects. Moreover, it could be argued that the Southgate et al. (2010) task is also demanding of inhibitory processes. In this task, infants need to ignore the location toward which the agent is pointing, and thus made most salient, and select the other box that the experimenter is ignoring.

⁴ It is possible that, if understood, the initial phrase "do you remember what I put in here?" could have cued infants to think about which object the experimenter put in there, rather than about what the experimenter thinks is in there. Thus, in an additional experiment, we used a different phrase ("do you know what is in here") with identical results.

⁵ As opposed to a gaze-based (looking-time or anticipatory-looking) response, which many have termed an "implicit" response. However, it should be recognized that this is simply a description of the two kinds of responses and does not tell us anything about the computations involved.

Thus, presumably, some degree of response-inhibition is required for this task, and there seems no *a priori* reason to assume that, if response inhibition is required to pass the standard Sally-Anne task, it is greater than would be required to pass this kind of “indirect” elicited-response task (see also Buttelmann Carpenter, & Tomasello, 2009).⁶

Thinking about what is involved in passing false-belief tasks

Evidence reviewed in previous sections should give us reason to doubt that the gap between infants passing non-verbal theory of mind tasks, and the age at which children pass standard theory of mind tasks can be explained either by appealing to behavioral rules (Perner & Ruffman, 2005), or the elimination of response selection and inhibition (Baillargeon et al., 2010). In the absence of a convincing explanation for why infants pass non-verbal false-belief tasks, but older children still fail verbal false-belief tasks, we should think carefully about just what is involved in success on these different tasks.

According to Scott, Baillargeon and colleagues’ response account, infants possess a decoupling mechanism that enables them “to hold in mind two distinct versions of a scene: one that corresponds to reality, and one that incorporates the agent’s false belief” (Scott et al., 2009). Furthermore, they argue that infants realize that other’s beliefs may not mirror reality (Onishi & Baillargeon, 2005). While they argue that response inhibition is not required for non-verbal false-belief task success, others disagree and argue that the lack of a bias toward responding based on the reality-based representation is puzzling given that (a) infants have notoriously poor inhibitory capacities and (b) even adults default to a reality-based interpretation of others’ behavior under certain circumstances (Ruffman & Perner, 2005; Samson & Apperly, 2010).

While it is often assumed that success on false-belief tasks (verbal or non-verbal) requires that children realize that others’ beliefs are distinct from their own (e.g. Carpendale & Lewis, 2006; Onishi & Baillargeon, 2005), this is not necessarily so. In early discussions of what constitutes an understanding of other minds, the emphasis was on representing other’s representations (Dennett, 1978) without emphasis of the need to appreciate the fact that other’s representations might conflict with reality. The development of the false-belief task evolved not out of a need to demonstrate any appreciation that beliefs can be false, but rather because predictions based on beliefs that are true, while they might reflect mental state reasoning, might just as easily reflect infant’s expectation that others will behave in accord with reality. To borrow Dennett’s Punch and Judy example, the fact that 4-year-olds squeal in anticipatory delight when they see Punch about to push the box (that he thinks contains Judy, even though the children have seen her escape) off the cliff suggests they were able to reason that he is acting on a mistaken belief that Judy is still in the box. They laugh because they know that Judy is not in the box, but that Punch thinks she is. It is only funny because children know both of these things: they know that Punch has a belief, and *that* his belief is false.

Do infants need to appreciate both of these things in order to generate a prediction concerning where an agent will search when she hasn’t seen the displacement of her desired toy? The answer is no, not necessarily. In order to generate a correct prediction about where someone will search for an object, infants need only to consider the agent’s representation of the object’s location. They do not need to understand or appreciate that her representation is false, or think about how it compares with reality. To illustrate, in Southgate et al. (2007), infants observe an agent turn around after a puppet has put a toy in the left-hand container. During the time she is oriented away from the scene, the

⁶ However, as we will discuss in the following section, inhibition may be required to a greater extent on tasks in which the infants own perspective is explicitly highlighted.

puppet returns to take the toy away. To correctly anticipate that the agent will search in the left-hand box upon her return, infants might reason that she will search in the left-hand box because she last saw it there even though it is actually now in the right-hand box. Alternatively, they might reason that she will search in the left-hand box because she last saw it there. The exact same prediction would be made irrespective of whether or not infants encode her representation as false or not.

Is it plausible to think that infants might represent the agent's representation, without considering how that representation compares with reality? One argument might be that, due to their limited ability to act on the world themselves, the most important events to encode and retain are those that are relevant for others' actions. The location of the ball as represented by the agent will be predictive of her actions, but the location of the ball as represented by the infant will not be. One possibility then is that infants maintain in working memory only the most relevant representation for their purposes. In non-verbal false-belief tasks, during habituation or familiarization trials, infants are essentially told that, what is relevant is what the agent will do. For example, in Southgate et al. (2007), infants observe two familiarization trials during which they are trained that when the windows light up and they hear this simultaneous chime sound, the agent will reach through one of the windows. Eliciting anticipatory saccades to the windows works only because the infant is motivated to care about this aspect of the event (either because they care anyway, or because we train them that this is what is relevant). Plenty of evidence shows that we do not always encode or retain information that we do not deem to be relevant, and that we pay attention to some things at the expense of others (e.g. Duncan, 2006). Although phenomena such as change or inattention blindness have been most extensively studied in adults, there is also evidence showing that infants encode or retain different aspects of events depending on what they perceive as being most relevant (e.g. Mareshal & Johnson, 2000; Yoon et al., 2008). Thus, it is not implausible to take seriously the possibility that infants might not encode, or might not retain, the true location of an object that is the target of someone else's actions.

Moreover, recent research suggests that sometimes, we may not store our own and others' perspectives as distinct versions of reality that we can keep separate and compare, and that our encoding of others' experiences can even interfere with our encoding of our own experience of an event. For example, in a task in which adults had to make rapid judgments about how many objects were present in a scene, Samson and colleagues found that reaction times were slower when the number of objects that were visible to the subject and an avatar were different, suggesting that the agent was also computing the scene from the others' perspective and had to overcome this representation in order to make a correct judgment (Samson, Apperly, Braithwaite & Andrews, 2010). An interesting aspect of these results was that adults were actually slightly faster to judge someone else's perspective than they were to judge their own perspective, raising the possibility that sometimes, someone else's perspective might be given greater importance than one's own. In a similar study, Kovacs and colleagues gave subjects an object-detection task (Kovacs, Teglas & Endress, 2010). Participants watched as a Smurf character rolled a ball behind an occluder. Sometimes the participant's and the agent's beliefs differed (i.e. the participant had seen the ball roll away while the agent was away, so the participant knew the object was absent, but the agent should think it is present) and sometimes they were the same (i.e. both saw the ball leave, or both saw the ball roll back behind the occluder). Kovacs et al. (2010) measured participant's reaction times to detecting the presence of the ball once the occluder was lowered and found that even when participants had seen the ball roll away, if the agent should think the ball is present (because he was not there when the ball rolled away), participants were just as fast at detecting the object as when they themselves should think it is behind the occluder (because they had not seen it roll away). Both these results suggest that sometimes we may encode the others' perspective even at the expense of our own. Kovacs and colleagues (2010)

also extended their paradigm to 7-month-old infants and found a similar phenomenon: infants who had observed a ball roll away (so should represent no ball behind the occluder), but who knew that the Smurf was absent when the ball rolled away (so the Smurf should represent the ball as still being behind the occluder), looked longer when the ball was revealed to be absent than when the infant and the Smurf had both seen the ball roll away. These results raise the possibility that infants may not be representing the other's perspective of an event as a separate version of reality that they could compare with their own, and that their own representation of the event does not take precedence. However, although these results tell us that infants encode the others' perspective, and that it interferes with their own representation of an event, they do not tell us whether sometimes infants might *preferentially* encode or retain the perspective of the other.

Nevertheless, this hypothesis would provide an alternative explanation for why infant's performance on false-belief tasks is not subject to an egocentric bias. If infants encode the other's perspective, but either fail to encode, or fail to maintain in memory, events that occurred during her absence (which would not alter, or be relevant to, her subsequent actions), then there would be no alternative reality to inhibit. Infants would preferentially retain the aspects of the event that are directly relevant for the other's expected action, and while this would constitute a metarepresentation, it would not amount to an understanding of belief as a concept which can be true or false, and the truth value of the others representation may not be considered.⁷ This kind of representation would suffice for infants to generate accurate predictions about what others will do, but it would also have limitations. Returning to Dennett's Punch and Judy example, infants might be able to understand that Punch wants to get rid of Judy and that Punch represents Judy as being in the box and so will push the box over the cliff. However, they would not find it funny in the way that 4-year-olds do because the humor of this event comes from the fact that children know that she's not really in the box, and they can compare their knowledge with Punch's belief, and it is this difference in representations that makes the event funny. Furthermore, while representing the others' perspective without considering its truth value would be sufficient for action predictions, it is likely to be insufficient for explaining why others are acting as they are. This kind of perspective-taking might enable you to predict where Sally will search, but you will not be able to explain why Sally comes up empty-handed (see Andrews, 2003 for a similar distinction).

Finally, if infants could pass false-belief tasks without understanding that the belief is false, it may provide an explanation for why children do not pass the standard false-belief task until around 4 years of age. In most non-verbal tasks, there is nothing that forces infants to consider the real state of affairs, or to consider the others' perspective as distinct from their own. Infants are free to attend to whatever they deem most relevant. On the other hand, in verbal false-belief tasks, instructions often make it explicit that the other person has a perspective that differs from the child's own perspective and experimenters typically highlight the reality by asking children memory questions like "where is the ball really?" or "what is in the box?". This kind of questioning may lead children to consider the real location of the object and, until they have the ability to deal appropriately with two conflicting representations, they may default to answering in terms of their own perspective.

Conclusions

Our view of infant's social cognitive abilities has undergone a radical transformation in the last decade with the publication of results suggesting that even very young infants are considering others' perspectives on the world, and are not, as was once thought, trapped in an egocentric viewpoint. The results of the non-verbal, and indirect elicited response, false-belief tasks suggest

⁷ Samson & Apperly advocate a similar position.

that, long before children turn four, they understand that others' own experiences modulate their actions. The discovery that this understanding exists early in childhood not only gives support to previous interpretations of children's behavior in naturalistic settings (Dunn, 1991), but also provides a solution to the tension that has existed between the views that young children are competent communicators but incompetent mindreaders (Breheny, 2006).

However, while infants can track events from others' perspectives, this need not imply that the representations that underlie their mindreading abilities are the same as the representations that support belief reasoning in older children and adults. For example, theory of mind comprises not only the formation of a representation of someone's thought or perspective, but the process of using that representation to generate predictions about how those thoughts will influence behavior (Dennett, 1978). However, while looking-time data tell us that 7- and 10-month-olds are representing events from others' perspectives (Kovacs et al., 2010; Luo, 2011), they do not tell us whether these younger infants can use these inferences to generate predictions about how others will behave. Furthermore, we do not know whether infants, like adults, understand that a belief can be false, or when attributing a belief to someone, whether young infants are considering the truth value of that belief. Finally, while we readily talk about infants' understanding of mental state concepts like "knowledge" and "beliefs," we do not really know what infants mental state concepts might look like. Do infants possess distinct mental state concepts like knowledge and belief? It is possible that initially, infants may have available only a core epistemic-state concept (P. Carruthers, personal communication), which only gradually becomes differentiated with development. The kind of nuanced mental state concepts that adults possess are likely to be culture- and language-dependent and, while there may exist core psychological concepts that are universal and available pre-linguistically, further mental state concepts may be dependent on linguistic input (Scholl & Leslie, 1999; Wellman, 1998). Thus, while the evidence suggests that infants are tracking others' epistemic states from an early age, it is important to recognize that there are various ways in which this understanding might differ from an adult-like understanding, while ensuring that infants are very good at figuring out what others will do.

Acknowledgements

I would like to thank Teodora Gliga for valuable discussion and comments on an earlier version of this chapter. VS is supported by a Wellcome Trust Research Career Development Fellowship (088427/Z/09/Z).

References

- Andrews, K. (2003). Knowing mental states: The asymmetry of psychological prediction and explanation. In Q. Smith and A. Jokic (Eds), *Consciousness: New Philosophical Perspectives* (pp. 201–219). Oxford: Oxford University Press.
- Baillargeon, R., Scott, R., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14, 110–18.
- Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, 20, 191–208.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition*, 21, 37–46.
- Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.
- Breheny, R. (2006). Communication and folk psychology. *Mind and Language*, 21(1), 74–107.

- Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-olds show false belief understanding in an active helping paradigm. *Cognition*, 112, 337–42.
- 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, 672–91.
- Carpendale, J., & Lewis, C. (2006). *How Children Develop Social Understanding*. Oxford: Blackwell.
- Dennett, D. C. (1978). Beliefs about beliefs. *Brain and Behavioral Sciences*, 1, 568–70.
- Duncan, J. (2006). Brain mechanisms of attention. *Quarterly Journal of Experimental Psychology*, 59, 2–27.
- Dunn, J. (1991). Understanding others: Evidence from naturalistic studies of children. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 51–61). Oxford: Blackwell.
- Friedman, O., & Petrashek, A. R. (2009). Children do not follow the rule “ignorance means getting it wrong”. *Journal of Experimental Child Psychology*, 102, 114–21.
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, 7(7), 287–92.
- Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56, 165–93.
- German, T. P., & Leslie, A. M. (2000). Attending to and learning about mental states. In P. Mitchell & K. Riggs (Eds), *Children's Reasoning and the Mind* (pp. 229–252). Hove: Psychology Press.
- Gomez, J. C. (2008). The evolution of pretence: From intentional availability to intentional non-existence. *Mind & Language*, 23(5), 586–606.
- Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to understanding of false belief and the appearance-reality distinction. *Child Development*, 59(1), 26–37.
- Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Hedger, J. A., & Fabricius, W. V. (2011). True belief belies false belief: Recent findings of competence in infants and limitations in 5-year-olds, and implications for theory of mind development. *Review of Philosophy and Psychology*, 2(3), 429–447.
- Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, 21, 101–14.
- Knudsen, B., & Liszkowski, U. (2012a). Eighteen- and 24-month-old infants correct others in anticipation of action mistakes. *Developmental Science*, 15(1), 113–122.
- Knudsen, B., & Liszkowski, U. (2012b). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*, 17(6), 672–691.
- Kovacs, A. M., Teglas, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012), 1830–4.
- Leslie, A. M. (1994). ToMM, ToBy and Agency: Core architecture and domain specificity. In L. Hirschfeld and S. 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, 247–53.
- Low, J., & Wang, B. (2011). On the long road to mentalism in children's spontaneous false-belief understanding: are we there yet? *Review of Philosophy and Psychology*, 2(3), 411–28.
- Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition*, 121(3), 289–98.
- Mareschal, D. & Johnson, M. H. (2003). The “what” and “where” of object representations in infancy. *Cognition*, 88, 259–76.
- Meltzoff, A. M., & Brookes, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology*, 44(5), 1257–65.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308, 255–8.

- Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a "theory of mind". *Philosophical Transactions of the Royal Society, B*, 362, 731–44.
- Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5(2), 125–37.
- Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, 308, 214–16.
- Piaget, J. (1954). *The Construction of Reality in the Child*. New York: Basic.
- Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, 7(4), 157–60.
- Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 4, 515–26.
- Ruffman, T. (1996). Do children understand the mind by means of simulation or a theory? Evidence from their understanding of inference. *Mind & Language*, 11, 388–414.
- Ruffman, T., & Perner, J. (2005). Do infants really understand false belief? Response to Leslie. *Trends in Cognitive Sciences*, 9, 462–3.
- Ruffman, T., Slade, L., Redman, J. (2005). Young infants' expectations about hidden objects. *Cognition*, 97, B35–43.
- Ruffman, T., Taumoepeau, M., & Perkins, C. (2011). Statistical learning as a basis for social understanding in children. *British Journal of Developmental Psychology*, 30(1), 87–104.
- Samson, D., & Apperly, I. A. (2010). There is more to mind reading than having theory of mind concepts: New directions in theory of mind research. *Infant and Child Development*, 19, 443–54.
- Samson, D., Apperly, I. A., Braithwaite, J., & Andrews, B. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5), 1255–66.
- Scholl, B. J., & Leslie, A. M. (1999). Modularity, development and "theory of mind". *Mind & Language*, 14, 131–53.
- Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development*, 80, 1172–96.
- Scott, R. M., Baillargeon, R., Song, H., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology*, 61(4), 366–95.
- Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science*, 22, 878–80.
- Song, H., & Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology*, 44(6), 1789–95.
- Song, H., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition*, 109, 295–315.
- Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, 16, 907–912.
- Southgate, V., & Csibra, G. (2009). Inferring the outcome of an ongoing novel action at 13 months. *Developmental Psychology*, 45, 1794–8.
- Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18, 587–92.
- Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18(7), 580–6.
- Surian, L., & Leslie, A. M. (1999). Competence and performance in false belief understanding: A comparison of autistic and normal 3-year-old children. *British Journal of Developmental Psychology*, 17, 141–55.
- Trauble, B., Marinovic, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' belief? *Infancy*, 15, 434–44.

- Verfaillie, K., & Daems, A. (2002). Representing and anticipating human actions in vision. *Visual Cognition*, 9, 217–32.
- Wellman, H. M. (1998). Culture, variation and levels of analysis in folk psychologies. *Psychological Bulletin*, 124, 33–6.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development*, 72, 655–84.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in children's understanding of deception. *Cognition*, 13, 103–28.
- Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69, 1–34.
- Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, 22(2), 145–60.
- Yoon, J. M. D., Johnson, M. H., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 13690–5.
- Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief? *British Journal of Developmental Psychology*, 30(1), 156–71.

Learning about the mind from evidence: Children's development of intuitive theories of perception and personality

Andrew N. Meltzoff and Alison Gopnik

Where does our understanding of the mind come from? Different theoretical perspectives have different views on this question. Strong modularity and core knowledge theories (e.g. Leslie, 2005) propose that the essentials of our adult understanding of others are in place initially, and development involves relatively small changes in that knowledge around the edges. Strong “embodiment” and “resonance” theories (e.g. simulation and mirror-neuron based accounts of mindreading, such as Gordon, 1996 or Gallese & Goldman, 1998) also do not focus on developmental change and argue that our understanding of the mind is fundamentally not inductive. Rather than learning about the mind from evidence, both these views see our understanding as due to relatively automatic and specialized triggering or resonance processes. We “take on” the mental states of others or project our own experiences on to them—rather than inferring those states from evidence.

In contrast, “theory-theory” accounts (Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994) propose that our understanding of the mind, at least in large part, involves learning abstract causal structures from evidence—hence, the analogy to theory change in science—involving initial hypotheses, tests, and conceptual revision. In the past, these claims were largely made on the basis of the naturally occurring changes in children's understanding of the mind over time. Moreover, it was unclear just what kinds of learning mechanism would allow children to learn about a complex and invisible system like the mind so swiftly and effectively.

In the past 10 years, however, this has begun to change. First, there are results from training studies with young children, which show that providing evidence can lead to changes in children's understanding of the mind. For example, Amsterlaw and Wellman (2006) and Slaughter and Gopnik (1996) both showed that three-and-a-half-year-olds who received evidence about beliefs shifted to a new understanding of belief more quickly than those who did not. Importantly, this extended not only to their performance on the classic false-belief task, but to their understanding of related concepts like the appearance/reality distinction and the sources of beliefs. Interestingly, children showed this effect most clearly when they were asked to explain, rather than just describe the evidence. Moreover, naturally occurring variations in the availability of evidence children receive can change the timing of their belief understanding. For example, deaf children of hearing parents have markedly delayed false-belief understanding (see Gopnik & Wellman, 2012, for a review).

Even with these training effects, however, we might argue that the incoming evidence simply accelerates or delays a naturally occurring change. A more powerful demonstration of the role of evidence comes when we design experiments in which we systematically give infants or children different kinds of new evidence about a system and see what kinds of inferences they draw. This

has been the approach taken in both the statistical learning literature and the causal learning literature. When children are provided new patterns of evidence under experimental control, and the different patterns of evidence lead them to different conclusions, it seems more obvious that the evidence itself is doing the causal work. For the most part, however, this work has focused on children's learning of language (e.g. Kuhl, 2004; Saffran, Aslin, & Newport, 1996), physical properties of objects (Wu, Gopnik, Richardson, & Kirkham, 2011), or physical causal relations (e.g. Bonawitz, Lim, & Schulz, 2007; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Meltzoff, Waismeyer, & Gopnik, 2012; Sobel & Kirham, 2006), rather than on their psychological learning. The experiments we discuss in this chapter move beyond this to examine infants' and young children's developing understanding of other people's minds.

Probabilistic models and Bayesian learning

In parallel with the field accumulating new data, theoretical work over the last 10 years (e.g. Gopnik, 2012; Gopnik et al., 2004; Meltzoff, Kuhl, Movellan, & Sejnowski, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011) has shown increasingly that it is possible to specify more precisely and formally how children learn from evidence. In particular, within the framework of probabilistic models and Bayesian inference we can think of children's learning as a process of hypothesis testing and revision (Gopnik, 2012; Tenenbaum et al., 2011). Children use probabilistic models to generate structured hypotheses, then test and revise those theories in a systematic way based on evidence. Moreover, rather than simply generating a yes or no decision about whether a particular hypothesis is true, Bayesian inference considers multiple hypotheses and assigns probabilities to those hypotheses. Bayesian methods let you determine the probability of possibilities. The integration of prior knowledge and new evidence in Bayesian reasoning also gives Bayesian inference a characteristic combination of stability and flexibility—a learner will be reluctant to give up a strongly-confirmed hypothesis, but even the most entrenched idea can be rejected if enough counter-evidence accumulates.

Moreover, according to the theory-theory view, children often are not just learning particular causal relations but are also learning abstract generalizations about causal structure. In fact, empirical research has shown that children develop more abstract, framework knowledge over and above their specific causal knowledge. For example, children may know in general that actions are caused by beliefs and desires without being able to say exactly which beliefs and desires are involved in any particular case.

These broader generalizations are important in both scientific and intuitive theories. Philosophers of science refer to “over-hypotheses” (Goodman 1955), or “research programs” (Laudan 1977), or “paradigms” (Kuhn 1962) to capture these higher-order generalizations. Cognitive developmentalists have used the term “framework theories” (Carey 2009; Wellman 1990; Wellman & Gelman 1992). For example, in their framework theories, children assume there are different kinds of variables and causal structure in psychology vs. biology vs. physics. In fact, they often understand these abstract regularities before they understand specific causal relationships (e.g. Simons & Keil, 1995).

Some nativists argue that this must mean that the more abstract causal knowledge is innate. In contrast, constructivists, including Piaget and theory theorists, hold that this abstract causal knowledge could be learned. How could this be?

Griffiths and Tenenbaum (2007, 2009; Tenenbaum, et al., 2011), inspired by both philosophy of science and cognitive development, have formulated computational ways of representing and learning higher-order generalizations about causal structure. Following Gelman, Carlin, Stern,

& Rubin (2003), they call their approach hierarchical Bayesian modeling (HBM) or, sometimes, theory-based Bayesian modeling. The idea is to have meta-representations, that is, representations of the structure of particular causal hypotheses, and of the nature of the variables and relationships involved in those causal networks. These higher-level beliefs can constrain the more particular causal hypotheses. Moreover, these higher-level generalizations can themselves be learned. HBMs stack up hypotheses at different levels. The higher levels contain general principles that specify which hypotheses to entertain at the lower level.

Computational work on HBMs has shown that, at least normatively, hierarchical Bayesian learning can actually work. Higher-level framework theories can, indeed, be updated in a Bayesian way via evidence that contacts only lower level hypotheses. Griffiths and Tenenbaum (2007) provide several simple demonstrations; Kemp, Perfors, & Tenenbaum (2007) and Goodman, Ullman, & Tenenbaum (2011) provide more comprehensive and complex ones. These demonstrations show that it is possible, in principle, to learn to proceed at several levels at once—not just at the level of specific hypotheses, but also at the level of specific theories and, even more abstractly, at the framework theory level.

Probabilistic models in development

Probabilistic models were originally designed to be ideal rational accounts of how a scientist or a computer could best solve a learning problem. They also have attractions as theories of the learning mechanisms of cognitive development. One attraction is that, at least in principle, this kind of learning would allow children to gradually move from one structured hypothesis to another very different hypothesis based on patterns of evidence—children would not be restricted to making small tweaks to innate modules or to simply accumulating new data. The probabilistic nature of Bayesian inference also captures the often gradual and piecemeal way that development proceeds. At the same time, the generative power of structured models and hypotheses might help explain the abstract and general character of children's inferences.

In addition, the probabilistic models view gives us a new way to think about the innate bases of cognition. Rather than thinking about innate perceptual-cognitive structures as firm “constraints” on the kinds of knowledge that a human can develop, an innate “prior” might weigh certain hypotheses as more likely than others, but even these hypotheses could be overturned with sufficient counter-evidence. The work on hierarchical Bayesian learning (Griffiths & Tenenbaum, 2007) suggests that “priors” may not only take the form of specific hypotheses about particular causal relationships, but may involve broader “framework principles” about general theoretical categories and causal relations. These framework principles shape many more specific hypotheses, but they may themselves be overturned with sufficient counter-evidence.

Developmental changes in understanding the mind

We suggest that in terms of our understanding of the mind, a strong prior and innate “framework principle” is that our own mental states and those of others are likely to be similar. We can think of this as a Bayesian version of the “like-me” hypothesis that we have argued for in the past (Meltzoff, 2007, 2013; Meltzoff & Gopnik, 1993). This assumption shapes the human infants' early learning about the mind, allowing a framework for preferring some hypotheses to others. It is, however, only the beginning of our learning about the mind. Within the framework principle, we can use evidence to elaborate on our initial understanding in complex and abstract ways. Eventually, with accumulating evidence concerning differences between our own perceptions, desires, and beliefs, and those of others, we can revise or overturn that framework principle, as shown by developmental research (e.g. Gopnik & Wellman, 2012; Moll & Meltzoff, 2011; Repacholi & Gopnik, 1997).

In this chapter, we take the problem of *developmental change* in children's understanding of the mind to be central, and illustrate the foregoing ideas with two examples. The field's (over) concern with the shift in children's verbal reasoning about false belief at 3–4 years of age has obscured the important fact that the human intuition of how the mind works is an extended process, including significant changes both much earlier and later than the classic preschool shift. We consider both an earlier and a later set of developments. In both cases, we show that providing children with particular patterns of evidence, whether evidence from their own experience or about the behavior of others, can lead to novel and systematic new causal models of how the mind works. Moreover, in both cases we invoke the idea of a Bayesian framework principle. The first example concerns infants' early understanding of other people's visual perception. We suggest that the initial framework principle adopted preverbally, and perhaps present at birth, constrains inferences, but is itself influenced by evidence. In the second example, the development of an understanding of personality traits, we suggest that this higher-order principle is actually initially inferred from data, but then acts as a constraint on further inferences.

Understanding perception

Recent studies show that infants use first-person visual experiences as evidence for a new understanding of the perceptions of others. The research is built from the finding that young infants make a puzzling error. In gaze-following studies, 12-month-olds follow a person's line of regard to an external object even when a blindfold occludes that person's viewpoint. They do not make this error, however, when the person closes his eyes (Brooks & Meltzoff, 2002, 2005). Why do young infants seem to have a privileged understanding of eye closing over and above other blindfolds?

One idea is that infants have extensive evidence about the causal relation between eye closure and visual experience, but initially have much less evidence about other kinds of occlusion. Eye closure is a biological motion with which infants have extensive first-person, agentive experience. Even very young infants have strong evidence about the causal relation between whether their eyes are open or closed, and their visual experience. They can easily perform informal "tests" to assess this causal link—they can control their own vision by closing and opening their eyes. When they close their eyes, the visual world goes black, and when they open them the world pops back into view. Perhaps infants use this evidence, along with their initial "like-me" causal framework principle, to make the attribution about others' visual experiences. What applies to me, also applies to you.

This predicts that if infants are given systematic evidence that blindfolds block their own view, they should suddenly make different attributions to others. Meltzoff and Brooks (2008) tested this idea with 12-month-olds. Infants were randomly assigned to three experimental groups that differed only in the nature of the evidence provided to the infants. Infants in the key treatment group were given massive experience with blindfolds (see Figure 2.1). When the infant looked at a toy, the adult blocked the view with a blindfold. She then lowered it in a playful manner, only to repeat the cycle for the next toy the child fixed. Infants experienced that their own view was blocked, but they were given no training about the adult's viewpoint. A control group involved a cloth made from the same material as the blindfold, but with a small window cut out of the center. Infants in this control received the same protocol (controlling for cloth raising/lowering); however, they could peer through the windowed cloth. In a second control group, infants were familiarized with the opaque cloth while it was laying flat on the table.

At the end of training, all three groups were given a standard gaze-following test. Infants were confronted with a blindfolded adult who turned toward the distal objects. Infants who had received first-person training on the opaque blindfold responded in a completely different manner to the

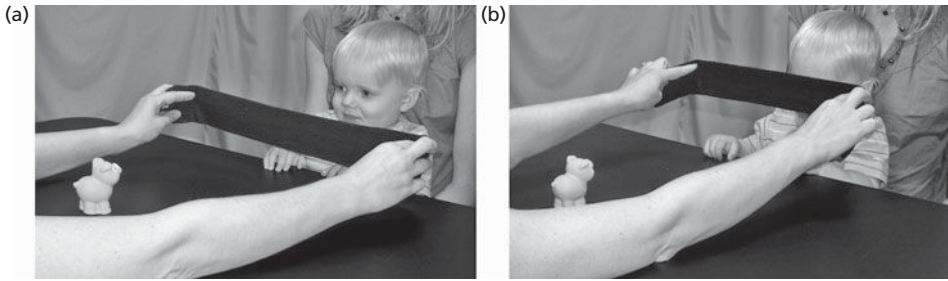


Figure 2.1 A 12-month-old boy in Meltzoff and Brooks' (2008) training procedure. Infants randomly assigned to a treatment group were given self-experience that a blindfold occluded their own perception. Infants looked at an interesting object (a). The blindfold then blocked their view of the toy (b). This was repeated over an 8-min training session. This first-person experience changed infants' interpretation of how opaque visual occluders influence other people's vision. See text for details.

Reproduced from Meltzoff, A.N. & Brooks, R. Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology*, 44, 1257–65. ©2008, American Psychological Association.

controls. Infants in this treatment group did not turn when the adult wore the blindfold, but infants in the controls still mistakenly followed the blindfolded adult's line of regard to the distal object, just like untreated infants in previous studies (e.g. Brooks & Meltzoff, 2002). It is as if infants in the treatment group had learned that the blindfold could not be seen through by *them*, and assumed that the same would be true for another person. They assumed the adult could not see. Therefore, there was no reason to follow his "gaze" when he turned to face the object; whatever the head turn was about, it was not a turn in order to see.

Making attributions about novel relations

One might argue that these experiments simply hastened a natural development—an understanding that you cannot see through opaque occluders. Could we use evidence similarly to teach infants a perceptual principle that they would not encounter naturally? Could that evidence even override a principle that they had learned earlier? In the natural course of development, by 18 months of age, infants no longer make the error of thinking that adults can see through opaque barriers (Brooks & Meltzoff, 2002; Butler, Caron, & Brooks, 2000; Dunphy-Lelii & Wellman, 2004). Meltzoff and Brooks (2008) capitalized on this to provide 18-month-olds with a completely novel self-experience—one they would not have encountered outside the laboratory. A trick blindfold was constructed that looked opaque from the outside, but was made of special material that could be seen through when held close to the eyes. Infants were randomly assigned to one of three groups: (a) experience with the trick blindfold, (b) experience with the opaque blindfold, and (c) baseline experience in which they played with the trick blindfold while it lay flat on a black table. After receiving the differential evidence, infants in all three groups saw the adult wear the blindfold in the standard gaze-following test.

As expected, infants in the baseline group and the opaque-blindfold groups refrained from following the adult's head turns when the adult wore the blindfold. The new finding is that infants who obtained evidence about the trick see-through blindfold now followed the adult's line of regard to the distal object—they treated the adult as if she could see despite wearing the opaque-looking occluder that covered her eyes.

This underscores the power of infant self-experience in making social attributions to others. Infants had learned that they could make perceptual contact with the external world through the blindfold. By employing the “like-me” framework principle, they immediately transferred this experience to others, despite the fact that the adult’s eyes were covered and it looked, from the outside, like she could not see. Moreover, this new experience now allowed them to override their earlier belief that blindfolds do obscure vision.

These results allow two inferences about development. The first is that infants are projecting their own inner experience to others, suggesting that by 12 months of age infants can attribute mental states (perceptual experience) to others. Crucially, the mentalism demonstrated is of an “on-off” variety, seeing vs. not seeing—a kind of perception-ignorance distinction. The current results do not show perspective-taking about *how* something appears to the other—only *that* it can be seen (or not) in the first place. It is widely argued that infants’ understanding of the basic on-off experience of vision is a building block for more complex mental states such as false belief. Of course, there are other findings suggesting that young children attribute visual experiences to others (e.g. Lempers, Flavell, & Flavell, 1977; Moll & Meltzoff, 2011; O’Neill, 1996; Onishi & Baillargeon, 2005; Repacholi, Meltzoff, & Olsen, 2008; Tomasello & Haberl, 2003). The specific advances of the current work are that it uses young infants (12-month-olds) and a controlled intervention paradigm with random assignment to show that infants use first-person evidence to *change* their understanding of the visual experiences of others.

A second inference concerns the level of abstraction at play. We believe that infants are learning about the spatial-causal relations among three entities: viewer, barrier, and object. These form a “visual perception triangle,” with the spatial relations determining whether the object can be seen by the viewer. Infants abstract a general lesson from the evidence of their own experience: “If the blindfold is interposed between viewer and object, the viewer cannot see the object.” This abstract description applies equally well to self and other. If infants can recognize that the spatial relation is similar—“blindfold over eyes”—they could generalize that the causal effect is similar.

The “like-me” causal framework principle allows infants to treat self and other as similar agents. What I learn about myself is immediately put to work in interpreting your behavior; reciprocally, the outcome of your actions on the world provides me with information about my own powers and the possibilities of my own future actions. This “like-me” framework principle is a human birthright (Meltzoff, 2007), underpinning unique features of human social learning and influencing the course of children’s development (Meltzoff, 2013; Meltzoff & Gopnik, 1993).

The “like-me” assumption supports learning about the world *from* watching other people. This occurs in cases of object-directed imitation and learning about cause and effect from observing social models (Meltzoff et al., 2012) as well as in learning abstract categorization rules from observing others’ sorting behavior (e.g. Williamson, Jaswal, & Meltzoff, 2010). The “like-me” assumption also supports learning *about* other people’s minds. Infants make attributions about the mental states of other “like-me” agents using their *own* first-person experience and mental states as a framework, which is a launching pad for developing an understanding of other minds.

Of course, philosophers have discussed whether an analogy between self and other plays a role in adult human affairs (e.g. Hume, 1739/1969; Husserl, 1950/1960; Smith, 1759/1966). The problem has traditionally been that the framework of equivalence was thought to be a late achievement and perhaps dependent on language, and therefore thought not to play a formative role during infancy. A quarter century of research on infancy has changed this view. In particular, the work on infant imitation indicates that young infants can represent the acts of others and their own acts in commensurate terms (Meltzoff & Moore, 1977, 1997). The generality of human imitation (face, hands, voice, object manipulation, styles of acting) establishes that human infants process a “like-me-ness”

at the level of behavior. They also recognize when their own acts are being reflected back or imitated by others, which prompts emotional and prosocial behavioral reactions by infants (e.g. Meltzoff, 2007) and special neural responses (Saby, Marshall, & Meltzoff, 2012).

What the blindfold training studies contribute is that the equivalence also is registered at the level of mental states. The infants in the blindfold gaze-following studies are not just registering equivalence in terms of behavior or visible configurations (e.g. “blindfold over eyes”), but inferring mental states. They are assuming that if a blindfold over their own eyes affects their vision, then it influences the *vision* of the blindfolded adult in the same way.

It is particularly striking how 18-month-olds react to experience with the trick see-through blindfold. Untreated children realize that opaque-looking occluders cannot be seen through, and do not follow the line of regard of a blindfolded adult. The novel intervention experience runs counter to everyday real-world experience. We arranged it so infants *can* see through this blindfold. Now when the adult dons the blindfold, infants interpret the behavior of the blindfolded adult in a new light. Now infants follow the blindfolded person’s “gaze” to distal objects. Infants attribute a psychological state (vision) to the blindfolded adult and interpret the adult’s behavior as a “turning to look.” In the absence of the novel self-experience, they do not do so.

In the cases we have described so far, evidence came from the child’s self-experience. Is self-experience the sole pathway to understanding others’ minds—a Royal Road? If we think of “like-me” as a framework principle then inferences should go in both directions—either from the self to the other or from the other to the self. We demonstrated just this in both 3-year-olds (Williamson & Meltzoff, 2011) and 18-month-olds (Meltzoff & Brooks, 2012).

In the latter study we arranged a situation in which 18-month-olds watched a blindfolded adult act in distinctive ways. The adult reached out and grabbed the toys, one by one, that were in front of her. To an adult it appeared that she was producing “visually-guided behavior.” It is as if the adult in this treatment group was demonstrating Superman’s X-ray vision. Control groups either performed the same behavior without a blindfold (controlling for “success” in grabbing the toys), or wore the blindfold and fumbled and missed the toys (controlling for “blindfold wearing”). After the exposure to the adult’s particular pattern of behavior (evidence accumulation), all infants were presented with the standard gaze-following test. Results showed that only the infants in the treatment group followed the gaze of the blindfolded adult. This suggests that self-experience is not the sole road for learning about other people’s minds. Infants can abstract information about whether the adult is (or is not) in visual contact with the world based on the cues, contingencies, and structural patterns that the other person exhibits while wearing the blindfold—that is, based on the patterning of *others’* behavior and not solely first-person experience.

To summarize, these experiments show that infants can combine an initial prior “like-me” framework principle with new evidence to infer new causal relations between objects, occluders, and experience both for themselves and others. These inferences go both ways—infants can make inferences about the behavior of others from their own experiences, but they can also make inferences about their own experiences from the behavior of others (e.g. that they will see something interesting if they follow the gaze of the “X-ray vision” adult).

We once suggested that the key thought experiment that would differentiate strong “modularity” theories from the “theory-theory” would be to place children in an alternative parallel universe with evidence that differs radically from our own (Gopnik & Meltzoff, 1997). If children developed a veridical understanding of that universe that would support the theory-theory; if they stuck to their innate understanding of this universe, that would favor the modularity predictions. However, we doubted if the granting agencies would have the funds to support the experiment. In these blindfold experiments, however, we have shown that we can do the same thing, although in a more

low-cost way. In effect, we presented children with alternative universes in which opaque-looking blindfolds are transparent, or in which some adults have the equivalent of Superman's X-ray vision. Even 18-month-olds made the correct inferences about what human behavior and experience would be like in this world.

Social attribution and the understanding of personality traits

If children are making new inferences about the mind from evidence well before they are three, they are also making new inferences well after they are five. These inferences are particularly interesting because they often straddle the unclear line between "theory of mind" and social psychology. One area of particular interest is the inferences we make about personality traits. A long tradition in social psychology (e.g. Kelley, 1967) shows that adults, at least in Western cultures, tend to explain people's actions in terms of their individual "personality traits." Our adult language is permeated with trait judgments, from brave to shy to intelligent to arrogant to introverted. Indeed, if I asked you what someone was like and you answered by giving me a description in terms of a 5-year-olds theory of mind ("well ... she believes that what she sees directly is true, and she usually tries to get what she wants ...") I would hardly be satisfied. Instead, I would expect some discussion of those personality traits that are consistent in her behavior and make her different from everyone else ("she is intelligent and charming but manipulative; he is difficult and bad-tempered but full of integrity"). This would allow me to predict her next move and explain to myself why the person acted toward me like she did.

Adults in Western societies tend to attribute behavior to such personality traits even when the evidence suggests that those actions are really the result of the situations people find themselves in. These attributions can, literally, be a matter of life and death. In the Abu Ghraib trials, for example, many observers initially attributed the atrocities to the sadistic individual personalities of those particular guards, despite the unsettling social psychology evidence suggesting that a wide range of people might behave equally badly in such circumstances.

Where do these attributions come from? It is unclear when and why children begin to explain action in terms of internal, individual, and enduring traits. Of course, even very young children tend to explain action in terms of internal mental states (Flavell, Green, & Flavell, 1990; Lillard & Flavell, 1990). However, trait explanations include two additional factors beyond mental states themselves. Traits are specific to particular, individual people, and they are constant over time and across situations.

Researchers have demonstrated that children do not spontaneously explain actions in terms of traits or endorse trait explanations for a single instance of behavior until middle childhood (Alvarez, Ruble, & Bolger, 2001; Peevers & Secord, 1973; Rholes & Ruble, 1984; Shimizu, 2000). However, other studies show that when preschoolers are given trait labels or behavioral frequency information, they can use that information to make inferences about future behavior, and that they can infer a trait label from frequent behaviors (Boseovski & Lee, 2006; Ferguson, Olthof, Luiten, & Rule, 1984; Gelman, 2003; Heyman & Gelman, 1999; Liu, Gelman, & Wellman, 2007; Matsunaga, 2002). On the other hand, these preschoolers still did not spontaneously construct trait explanations; rather they simply matched the frequency of behaviors to trait labels that were provided for them. This suggests that the failure to attribute traits is not simply a problem with word comprehension or conceptual resources, but may reflect something specific to the child's social cognition. Moreover, when children saw one behavior that could suggest a trait, e.g. one brave action, they did not predict that other behaviors would follow suit, as adults do.

We do not know the learning mechanisms that underlie the course of trait attributions in childhood. Kelley provided an early theory suggesting that person and situation covariation evidence might play an important role in attributions in adults (Kelley, 1967), and there is significant work in this area by social psychologists (e.g. Mischel & Shoda, 1995; Mischel, Shoda, & Mendoza-Denton, 2002; Plaks, Grant, & Dweck, 2005; Ross, 1977). Empirical studies confirm that adults use statistical information tracking multiple people in multiple situations to make trait attributions (e.g. Cheng & Novick, 1990; Hewstone & Jaspars, 1987; Morris & Larrick, 1995; Orvis, Cunningham, & Kelley, 1975; Sutton & McClure, 2001). However, adults already have intuitive theories of traits. They use covariation data to decide when and how to apply those theories to interpret and predict behavior, but could covariation play a role in the *development* of trait attribution itself?

Bayesian causal learning theories suggest that children systematically combine prior knowledge and current covariation evidence to arrive at the right causal hypothesis. This suggests a potential mechanism for the development of trait attribution. Children may begin by observing person and situation covariation evidence that confirms a particular type of hypothesis, particularly in conjunction with adults' linguistic accounting that internal traits cause actions. Once that theory has been strongly confirmed, it will be more difficult to overturn in the future, although it might still be overturned with sufficient evidence. Eventually, in adults raised in Western societies (Nisbett, 2003), this may result in a consistent "trait bias" that requires a very large amount of contrary evidence or concentrated effort to overcome (e.g. Blackwell, Trzesniewski, & Dweck, 2007; Dweck, 2006).

In a series of studies, Seiver, Gopnik, and Goodman (2013) examined the developmental origins of trait attribution in children raised in the USA. First, they conducted a study where 4- and 6-year-old children observed a scenario of two dolls playing on two activities (e.g. a bicycle and a trampoline). Children were either in the person condition (where the two doll characters acted consistently on the two activities, and differently from each other) or the situation condition (where both dolls played on one toy activity and did not play on the other). In some of the conditions this evidence was probabilistic—the doll would play on the toy either three out of four times or one out of four times. The children in each condition received different covariation information about the person and situation. In the person condition, covariation pattern of data indicated that some trait of the individual doll was responsible for the action; in the situation condition, the covariation pattern indicated that the situation was responsible. However, in both conditions, overall, there were the same number of examples of playing and not playing. At the end, we asked the children to explain the doll's actions (e.g. "Why did Josie play on the bicycle?") and to predict their behavior in a future situation (e.g. "What will Josie do when she sees this new diving board? Will she play on it or not?"). We also asked them to predict a new doll's response to the same situations (e.g. "What if Mary sees the trampoline, will she play on it or not?")

In the person condition, one doll always plays and the other doll never plays. This pattern of evidence suggests that something internal about the individual, rather than the situation is responsible for her behavior. In the situation condition, both characters never play with one toy and always play with the other, suggesting instead, that the situation or the toy itself is responsible for their actions. So how would children explain the dolls' behavior in these two different conditions?

Four-year-olds offered explanations that matched the pattern of evidence. In the person condition, when the evidence indicated that something about the person was responsible for the dolls' behavior, both 4- and 6-year-olds gave internal explanations for that behavior. Interestingly, and in keeping with earlier findings, these were rarely classic trait explanations, especially for the 4-year-olds. Instead, children offered explanations that highlighted "trait-like" characteristics of the person, which included both physical characteristics like age or height ("she's the big sister,"

“she’s only little”) and mental states, such as long-standing desires and beliefs (“she likes playing on bicycles,” “she thinks the water is dangerous”).

In the other condition, when the evidence suggested that the situations were driving the dolls’ actions (i.e. they both played on one activity and did not play on the other), 4-year-olds also appropriately gave more external explanations—explanations involving the specific toy activity (e.g. that the bicycle was tippy or the trampoline was safe). In contrast, 6-year-olds persisted in giving internal explanations. Like the Western adults interpreting the events of Abu Ghraib, they attributed the dolls actions to their internal states, even when the pattern of evidence went against those attributions.

In fact, this difference in attribution style between the two age groups in the situation condition suggests that the 4-year-olds were actually more sensitive to the covariation data than the 6-year-olds—they were actually better or, at least, more open-minded learners given the pattern of evidence. Seiver et al. (2013) also included a control condition where children were asked to explain why a single doll did or did not play on a single activity. In this case, the pattern of evidence provided to the children was ambiguous about the possible cause of the behavior. Six-year-olds gave internal explanations significantly more often than expected by chance; 4-year-olds were at chance. The prediction question provided additional evidence for the same developmental change.

This pattern of results suggests that American 6-year-olds have developed a specific attributional theory or person “schema”—that is a broad framework principle—that the internal qualities of a person, rather than the situation, drives behavior. This existing framework principle acted as a filter on their interpretation of the data favoring trait explanations. It did this in much the same way that infants’ “like-me” framework drove them to immediately generalize their own experience to those of others. While we argue that the “like-me” principle has an innate foundation, the trait framework seems to be something that children learn in Western society (best estimate is about 6 years of age). Six-year-olds use both the evidence at hand and their prior beliefs to arrive at a conclusion about a person-situation scenario. The 4-year-olds in contrast, seem to use a more general “bottom-up” data based strategy, and only use the most immediately available data to draw conclusions about other people’s personality.

Domains and development

An interesting characteristic of hierarchical Bayesian learning is that broad framework principles can actually constitute domains. That is, when children learn a new overarching principle that applies to a particular set of data, that principle can act as a constraint on their further inferences. Some principles, like the “like-me” principle could already divide up the world into domains very early. Indeed, there is reason to believe that young infants divide the world into “like-me” and “not-like-me” domains—at a first approximation, animate experiencing agents and inanimate unconscious objects—and treat those domains as if they follow separate rules. However, other principles like the “trait bias” could be learned from cultural-linguistic input and yet have the similar far-reaching effects overall. We can also ask how domain-specific or how general this higher-order bias actually is. Does it only apply to the case of psychological causation, or would children reason similarly about internal versus external causes of physical outcomes?

From people to magnets

To explore potential attributional bias in understanding physical causation (Seiver et al., 2013) changed the outcome of interest to a physical, rather than psychological one—“stickiness” instead of willingness to play. Without changing the task in any other way, they altered the cover story

to implicate physical instead of psychological causation. Rather than saying that the doll character was playing on the scooter, we would say that the doll was sticking to the scooter. The relevant explanatory question became, “Why did the Josie doll stick to the scooter?” For “internal” responses children talked about the properties of the doll; for “external” responses they talked about properties of the toy.

Four-year-olds in this condition behaved as they did in the psychological case. They continued to give more internal explanations in the doll condition, and external explanations in the toy condition. So 4-year-olds seemed to rely on the data, rather than on prior framework principles. However, 6-year-olds behaved differently: They lost their overall preference for internal explanations. Moreover, 6-year-olds now reliably extended the data pattern in both conditions to make future predictions. (Six-year-olds were still less likely to normatively explain the data than the 4-year-olds, however.)

Closer examination of the results suggests interesting details about the 6-year-olds’ shift from largely relying on the pattern of data provided to relying on a prior framework principle. In the physical case, the 6-year-olds gave explanations in terms of a different everyday causal theory—namely, magnetism. They appealed to the properties of magnetism, such as the relationship between magnets and metal, in their explanations and were more likely to give interactive causal explanations that implicated both the doll and the toy as causes for the outcome (e.g. “she has metal shoes and the skateboard is a magnet”). Children never produced these interactive explanations in the social case, and 4-year-olds rarely produced them in the physical case. These explanations suggest that the 6-year-old children relied on a more culturally-conferred, scientifically-based causal framework about stickiness and magnetism in particular, rather than relying on the pattern of observed data *per se*.

What kinds of evidence could lead to this developmental change? One interesting hypothesis is that the developments at about 6-years of age are related to the increase in peer group interaction. In many peer interactions in the USA, individual traits, rather than social roles or situations, will account for much of the variance in behavior. In a classroom of 28 otherwise similar children placed in a similar situation on the playground, some will consistently take risks and others will not. Children will see more trait-based covariation as they pay increasing attention to their peers, and acquire rich data sets across individuals and situations to draw upon.

Similarly, cross-cultural differences in covariation evidence may influence the development of attribution (Nisbett, 2003). Miller (1984) suggested that children across cultures began with similar attribution patterns and then diverged toward the more extreme adult patterns as they grew older, a claim which has been supported by further studies with children (Gonzalez, Zosuls, & Ruble, 2010; Kalish, 2002; Lockhart, Nakashima, Inagaki, & Keil, 2009).

These results suggest a mechanism by which cultural differences may influence the course of social attribution. This may either be because members of different cultures actually do behave differently, or because culture and experience influence the information children receive from adults about traits, such as adult trait language. This evidence is especially relevant to the development of person schemata. If the adults within a culture tend to linguistically describe and label behavior in terms of traits, this will lead to covariation between certain behaviors and trait labels, which might itself provide evidence for a trait-schema (see Kemp, Goodman, & Tenenbaum, 2008). If children are using covariation information about people’s behavior *and* adult trait language to make inferences about people in situations, such differences in the data could affect the development of their mature adult social cognition. An interesting test would be to explore children in a less trait-based culture, e.g. mainland China. One might predict that 4-year-olds would show a similar pattern to what we observed, but 6-year-olds would not manifest the same trait bias.

Conclusion

We have provided two examples about infants' and children's developing understanding of other minds, one substantially before the well-researched 3–4-year-old age period, and one after it. The first example concerned infants' early attribution of mental states (visual perception) to others, and the second example concerned children's changing interpretation of others' personalities. In both cases, we claim that infants and children are using evidence to develop new inferences and models of other people's minds. These inferences can be specific causal hypotheses about what will happen in a particular situation—the adult will see the distal objects through the opaque-looking occluder or not; the doll will play on one toy, rather than another. They may also, however, involve inferences from and about general framework principles. Another person will experience the world in the same way that I do. People act based on their individual traits, rather than the situations they find themselves in. When we systematically manipulate the evidence that children receive, they draw different conclusions about the nature of the minds of those around them.

In the real world, children within a particular cultural milieu (shared language, customs, and physical world) may receive reasonably consistent, statistically discernable patterns of evidence about some aspects of the mind, such as visual perception, and so converge on the same general theories as other members in their culture. However, the example of traits and others provided in this chapter also emphasize that many aspects of mental life are likely to vary in different places and different times, and in the myriad of social, physical, and virtual environments that human beings create. Powerful theory-like inferential abilities may be particularly valuable in that sort of world.

One of the most endearing and powerful aspects about the child's social mind is that they change it based on evidence. Adult theorists are challenged to create theories explaining children's conceptual plasticity and developmental trajectory.

Acknowledgement

Work on writing this chapter was supported by NSF grants SMA-0835854 (ANM) and BCS-1023875 (AG). NSF is not responsible for the content. We thank participants in the McDonnell Causal Learning Collaborative for discussions on these topics. The authors contributed equally to this chapter.

References

- Alvarez, J.M., Ruble, D.N., & Bolger, N. (2001). Trait understanding or evaluative reasoning? An analysis of children's behavioral predictions. *Child Development*, 72, 1409–25.
- Amsterlaw, J., & Wellman, H.M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, 7, 139–72.
- Blackwell, L.S., Trzemieski, K.H., & Dweck, C.S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, 78, 246–63.
- Bonawitz, E.B., Lim, S., & Schulz, L.E. (2007). Weighing the evidence: Children's naïve theories of balance affect their exploratory play. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 113–18). New York, NY: Erlbaum.
- Boseovski, J.J., & Lee, K. (2006). Children's use of frequency information for trait categorization and behavioral prediction. *Developmental Psychology*, 42, 500–13.
- Brooks, R., & Meltzoff, A.N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, 38, 958–66.

- Brooks, R., & Meltzoff, A.N. (2005). The development of gaze following and its relation to language. *Developmental Science*, 8, 535–43.
- Butler, S.C., Caron, A.J., & Brooks, R. (2000). Infant understanding of the referential nature of looking. *Journal of Cognition and Development*, 1, 359–77.
- Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.
- Cheng, P.W., & Novick, L.R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality & Social Psychology*, 58, 545–67.
- Dunphy-Lelii, S., & Wellman, H.M. (2004). Infants' understanding of occlusion of others' line-of-sight: Implications for an emerging theory of mind. *European Journal of Developmental Psychology*, 1, 49–66.
- Dweck, C. (2006). *Mindset: The New Psychology of Success*. New York: Random House.
- Ferguson, T.J., Olthof, T., Luiten, A., & Rule, B.G. (1984). Children's use of observed behavioral frequency versus behavioral covariation in ascribing dispositions to others. *Child Development*, 55, 2094–105.
- Flavell, J.H., Green, F.L., & Flavell, E.R. (1990). Developmental changes in young children's knowledge about the mind. *Cognitive Development*, 5, 1–27.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2, 493–501.
- Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2003). *Bayesian Data Analysis* (2nd ed.). New York: Chapman & Hall.
- Gelman, S.A. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford: Oxford University Press.
- Gonzalez, C.M., Zosuls, K.M., & Ruble, D.N. (2010). Traits as dimensions or categories? Developmental change in the understanding of trait terms. *Developmental Psychology*, 46, 1078–88.
- Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge: Harvard University Press.
- Goodman, N.D., Ullman, T.D., & Tenenbaum, J.B. (2011). Learning a theory of causality. *Psychological Review*, 118, 110–19.
- Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, 337, 1623–7.
- Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 3–32.
- Gopnik, A., & Meltzoff, A. (1997). *Words, Thoughts, and Theories*. Cambridge: MIT Press.
- Gopnik, A., & Wellman, H. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 257–93). New York: Cambridge University Press.
- Gopnik, A., & Wellman, H.M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, 138, 1085–108.
- Gordon, R.M. (1996). “Radical” simulation. In P. Carruthers & P. K. Smith (Eds), *Theories of Theories of Mind* (pp. 11–21). Cambridge: Cambridge University Press.
- Griffiths, T. L., & Tenenbaum, J.B. (2007). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds), *Causal Learning: Psychology, Philosophy, and Computation* (pp. 323–345). New York: Oxford University Press.
- Griffiths, T.L., & Tenenbaum, J.B. (2009). Theory-based causal induction. *Psychological Review*, 116, 661–716.
- Hewstone, M., & Jaspars, J. (1987). Covariation and causal attribution: A logical model of the intuitive analysis of variance. *Journal of Personality and Social Psychology*, 53, 663–72.
- Heyman, G.D., & Gelman, S.A. (1999). The use of trait labels in making psychological inferences. *Child Development*, 70, 604–19.
- Hume, D. (1969). *A Treatise of Human Nature*. London: Penguin Books (original work published 1739).

- Husserl, E. (1960). *Cartesian Meditations: An Introduction to Phenomenology* (D. Cairns, trans.). The Hague: Nijhoff (original work published 1950).
- Kalish, C.W. (2002). Children's predictions of consistency in people's actions. *Cognition*, **84**, 237–65.
- 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.
- Kemp, C., Goodman, N.D. & Tenenbaum, J.B. (2008). Theory acquisition and the language of thought. Paper presented at the 30th Annual Conference of the Cognitive Science Society, Washington, D.C..
- Kemp, C., Perfors, A., & Tenenbaum, J.B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, **10**, 307–21.
- Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, **5**, 831–43.
- Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Laudan, L. (1977). *Progress and its Problems: Toward a Theory of Scientific Growth*. Berkeley: University of California Press.
- Lempers, J.D., Flavell, E.R., & Flavell, J.H. (1977). The development in very young children of tacit knowledge concerning visual perception. *Genetic Psychology Monographs*, **95**, 3–53.
- Leslie, A.M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, **9**, 459–62.
- Lillard, A.S., & Flavell, J.H. (1990). Young children's preference for mental state versus behavioral descriptions of human action. *Child Development*, **61**, 731–41.
- Liu, D., Gelman, S.A., & Wellman, H.M. (2007). Components of young children's trait understanding: Behavior-to-trait inferences and trait-to-behavior predictions. *Child Development*, **78**, 1543–58.
- Lockhart, K.L., Nakashima, N., Inagaki, K., & Keil, F.C. (2009). From ugly duckling to swan? Japanese and American beliefs about the stability and origins of traits. *Cognitive Development*, **23**, 155–79.
- Matsunaga, A. (2002). Preschool children's inferences about personality traits. *Japanese Journal of Developmental Psychology*, **13**, 168–77.
- Meltzoff, A.N. (2007). 'Like me': A foundation for social social cognition. *Developmental Science*, **10**, 126–34.
- Meltzoff, A.N. (2013). Origins of social cognition: Bidirectional self-other mapping and the "Like-Me" hypothesis. In M. Banaji & S. Gelman (Eds), *Navigating the Social World: What Infants, Children, and Other Species Can Teach Us* (pp. 139–44). New York: Oxford University Press.
- Meltzoff, A.N., & Brooks, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology*, **44**, 1257–65.
- Meltzoff, A.N. & Brooks, R. (2012). *Infants inferring mental states from observing others' behavior*. Unpublished manuscript. Seattle: Institute for Learning & Brain Sciences. University of Washington.
- Meltzoff, A.N., & Gopnik, A. (1993). The role of imitation in understanding persons and developing a theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding Other Minds: Perspectives from Autism* (pp. 335–66). New York: Oxford University Press.
- Meltzoff, A.N., Kuhl, P.K., Movellan, J., & Sejnowski, T.J. (2009). Foundations for a new science of learning. *Science*, **325**, 284–8.
- Meltzoff, A.N., & Moore, M.K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, **198**, 75–8.
- Meltzoff, A.N., & Moore, M.K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting*, **6**, 179–92.
- Meltzoff, A.N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: Observational causal learning in 24-month-old infants. *Developmental Psychology*, **48**, 1215–28.
- Miller, J.G. (1984). Culture and the development of everyday social explanation. *Journal of Personality and Social Psychology*, **46**, 961–78.

- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–68.
- Mischel, W., Shoda, Y., & Mendoza-Denton, R. (2002). Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science*, *11*, 50–4.
- Moll, H., & Meltzoff, A.N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development*, *82*, 661–73.
- Morris, M. W. & Larrick, R. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*, 331–55.
- Nisbett, R.E. (2003). *The geography of thought: How Asians and Westerners think differently ... and why*. New York: Free Press.
- O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development*, *67*, 659–77.
- Onishi, K.H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–8.
- Orvis, B.R., Cunningham, J.D., & Kelley, H.H. (1975). A closer examination of causal inference: The roles of consensus, distinctiveness, and consistency information. *Journal of Personality and Social Psychology*, *32*, 605–16.
- Peevers, B.H., & Secord, P.F. (1973). Developmental changes in attribution of descriptive concepts to persons. *Journal of Personality and Social Psychology*, *27*, 120–8.
- Plaks, J.E., Grant, H., & Dweck, C. (2005). Violations of implicit theories and the sense of prediction and control: Implications for motivated person perception. *Journal of Personality and Social Psychology*, *88*, 245–62.
- Repacholi, B.M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month olds. *Developmental Psychology*, *33*, 12–21.
- Repacholi, B.M., Meltzoff, A.N., & Olsen, B. (2008). Infants' understanding of the link between visual perception and emotion: "If she can't see me doing it, she won't get angry." *Developmental Psychology*, *44*, 561–74.
- Rholes, W.S., & Ruble, D.N. (1984). Children's understanding of dispositional characteristics of others. *Child Development*, *55*, 550–60.
- Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in social psychology*, Vol. 10 (pp. 173–220). New York: Academic Press.
- Saby, J.N., Marshall, P.J., & Meltzoff, A.N. (2012). Neural correlates of being imitated: An EEG study in preverbal infants. *Social Neuroscience*, *7*, 650–61.
- Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–8.
- Seiver, E., Gopnik, A., & Goodman, N.D. (2013). Did she jump because she was the big sister or because the trampoline was safe? Causal inference and the development of social attribution. *Child Development*, *84*, 443–54.
- Shimizu, Y. (2000). Development of trait inference: Do young children understand the causal relation of trait, motive, and behavior? *Japanese Journal of Educational Psychology*, *48*, 255–66.
- Simons, D.J., & Keil, F.C. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition*, *56*, 129–63.
- Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: Training children to understand belief. *Child Development*, *67*, 2967–88.
- Smith, A. (1966). *The theory of moral sentiments*. New York: Augustus M. Kelley. (Original work published 1759).
- Sobel, D.M., & Kirkham, N.Z. (2006). Bickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, *42*, 1103–15.

- Sutton, R.M., & McClure, J. (2001). Covariational influences on goal-based explanation: An integrative model. *Journal of Personality and Social Psychology*, 80, 222–36.
- Tenenbaum, J.B., Kemp, C., Griffiths, T.L., & Goodman, N.D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331, 1279–85.
- Tomasello, M., & Haberl, K. (2003). Understanding attention: 12- and 18-month-olds know what is new for other persons. *Developmental Psychology*, 39, 906–12.
- Wellman, H.M. (1990). *The Child's Theory of Mind*. Cambridge: MIT Press.
- Wellman, H., & Gelman, S. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43, 337–75.
- Williamson, B.A., Jaswal, V.K., & Meltzoff, A.N. (2010). Learning the rules: Observation and imitation of a sorting strategy by 36-month-old children. *Developmental Psychology*, 46, 57–65.
- Williamson, R.A., Meltzoff, A.N. (2011). Own and others' prior experiences influence children's imitation of causal acts. *Cognitive Development*, 26, 260–8.
- Wu, R., Gopnik, A., Richardson, D.C., & Kirkham, N.Z. (2011). Infants learn about objects from statistics and people. *Developmental Psychology*, 47, 1220–9.

Chapter 3

Teleology: Belief as perspective

Johannes Roessler and Josef Perner

A fundamental question in recent “theory of mind” research is how to interpret a seemingly robust dissociation between young children’s performance on different kinds of tests for false belief understanding. 3-year-olds’ poor performance on classical, “direct” false belief tasks is well-documented. Yet a range of “indirect” tests reveal sensitivity to agents’ false beliefs in much younger children. It is natural to think that the two kinds of tests bring to light two kinds of understanding: “explicit” vs. “implicit” understanding. But how should we understand this distinction? And why should “implicit” understanding of false beliefs only be available in connection with “indirect” tests?

Our project in this chapter is to address these questions by further developing a hypothesis advanced elsewhere (Perner and Roessler, 2010). This is the hypothesis that young children are *teleologists*: they make sense of intentional actions in terms of justifying reasons provided by “worldly” facts (not by mental states). We begin by spelling out this account in more detail. We then argue that mastery of the concept of belief (or possession of an “explicit understanding” of belief) involves giving a twist to the teleological scheme of explanation. What is critical is the ability to engage in hypothetical or suppositional reasoning about justifying reasons. This account, we contend, is in competition with both a “theory theory” and a “simulation theory” of belief understanding (though it has some affinities with certain versions of the latter). In the final, fourth part of the chapter we bring the account to bear on the dissociation problem. The difference between “direct” and “indirect” tests, we argue, turns on whether successful performance requires understanding the normative underpinnings of the causal role of belief (as in direct tests) or merely requires a set of generalizations regarding the causes of behavior (as in indirect tests).

Teleological explanation

Why does the baker get up at 3 a.m.? Well, the bread needs to be ready by 6 to go to the supermarkets, and it takes that long to bake. This is a humble example of a teleological explanation: it makes the baker’s unusual behavior intelligible not by appeal to his mental states, such as his desire to make bread etc, but in terms of the objective reason-giving facts of his situation. Our suggestion is that young children are teleologists. They predict, and perhaps explain, what someone will do on the basis of what it makes objective sense for her to do. This, we suggest, explains the following striking finding concerning young children’s performance on false belief tests: far from answering the test question randomly, they systematically and adamantly give the wrong answer. The explanation is that they predict that the protagonist will do what he *ought to do* in order to attain his objective. For example, they will predict that in order to retrieve his chocolate Mistaken Max (Wimmer & Perner, 1983) will go to the cupboard (where he ought to go, as this is where the chocolate is to be found) rather than to the kitchen drawer (where he believes it is).

This needs some elaboration and qualification. You might say that there is a sense in which Max ought to go *to the drawer*. Given his false belief, surely it would be quite irrational for him to go the cupboard, where he has absolutely no reason to expect the chocolate. We agree. But the point is consistent with there *also* being a sense in which he ought to go to the cupboard: we have to recognize two kinds of practical “ought.” Sometimes we are interested in whether someone ought to perform a certain action in the sense that there is *reason* for her to perform it. (An obvious context in which this question is to the fore is when you deliberate about what to do—i.e. reflect on what you have reason to do.) Sometimes we are interested in whether an agent *rationally* ought to perform a certain action, given her existing beliefs, aims and dispositions. Following Kolodny (2005) we will call these the “ought” of reasons vs. the “ought” of rationality. (Alternatively, one might put the contrast in terms of objective vs. subjective reasons.)

To make the distinction vivid, suppose we are advising Mistaken Max on what to do. From our vantage point as spectators of the story, the obvious recommendation is: “You ought to go to the cupboard—you have reason to: that’s where the chocolate is.” On the other hand, if Max remains firmly convinced that the chocolate is in the drawer, despite our best efforts to convince him otherwise, we might switch to a different kind of advice: “Given your belief, the rational thing for you to do is clearly to go to the drawer—that’s what you ought to do.”¹ The first type of judgment is more prevalent in the context of advice and joint deliberation; the second type of judgment is more prevalent in the context of evaluating the rationality of an action. But they are both central and familiar elements of commonsense psychology.

Another way to bring out the distinction is to consider what happens when Mistaken Max (without the benefit of advice from us) goes to the kitchen table and opens the drawer. No doubt he’ll be surprised. That’s because he realizes that he was *wrong*: he thought there was a good reason for him to go to the drawer, but it now turns out that there wasn’t. No chocolate—no reason. This is of course consistent with saying, as Max may find it comforting to say, that it was perfectly *rational* for him to go to the drawer.

The “ought” of rationality is often invoked in the context of action *explanation*. Those who emphasize the “rationalizing” nature of such explanations tend to have in mind that we explain intentional actions in terms of attitudes—centrally, beliefs and desires—that make it rational for the agent to perform the action. Our proposal is that young children think of intentional actions in a more simple-minded way: they predict and explain actions in terms of *reason*-giving facts, rather than *rationalizing* mental states.

But can such facts coherently be conceived as *causes*? As Davidson taught us, to explain why someone got up at 3 a.m., it is not enough to assemble considerations—“justifying reasons”—that show this to have been the right thing for him to do. What is required is a causal explanation (Davidson 1963). We grant the point. But we suggest that there is nothing incoherent in the idea that reason-giving “worldly” facts causally explain someone’s actions. Note, first, that such facts yield reasons that can be “agent-specific.” That the bread needs to be ready by 6, happily, has no implications as to when *you* ought to get up, but it gives the baker and his staff a reason to rise early. This is not because the reason in question is provided by the beliefs and desires of the relevant agents. Rather, agent-specificity is secured in this schema either by dint of the social roles of the agents (it is the baker’s job to deliver the bread on time) or as a result of their practical abilities and opportunities for action (that Max needs his chocolate could give anyone a reason to help him get it if they are in a position to do so). You might say that without appeal to the agent’s beliefs and

¹ These pieces of advice are modeled on Kolodny’s examples of what he calls “objective” and “subjective” advice.

desires it's totally mysterious by what sorts of causal mechanisms the reason-giving facts impact on the agent's movements. But this does not impugn the *coherence* of the teleological schema, at least on what is sometimes called a "difference-making" approach to causal explanation (Woodward, 2011). To say that one fact causally explains another is to say that certain counterfactual conditionals hold: roughly, had there been some variation in respect of the first fact, there would have been a corresponding difference in the second fact. If the bread had not been needed until 7, the baker would have slept longer. You might still insist that, without some idea of the causal mechanisms involved, it would be quite irrational to make a causal judgment. Be that as it may (and the point is far from obvious), our claim is not that young children's simple-minded teleology is correct (or a model of rationality)—merely that it is a coherent explanatory (and indeed causal-explanatory) schema.

We have mentioned one piece of evidence in favor of the teleological analysis. If young children predict what people will do on the basis of teleological reasoning, it becomes comprehensible why they are so wedded to their predictions. It is not that they are unable to inhibit a prepotent response (in which case one would expect that once the mistake is pointed out, they realize what the correct answer is). Rather, their predictions are based on sound reasoning! People normally do what it makes sense for them to do. From a teleological point of view, what it makes sense for people to do depends on their objective circumstances—the relevant evaluative and instrumental facts. In other words, young children subscribe to a rather austere version of the "principle of charity," enjoining them to assume that people do what they have reason to do.²

Admittedly, the point hardly amounts to an open-and-shut case for teleology. Our aim here, though, is not to undertake a comprehensive review of the evidence. (See Perner and Roessler, 2010, for more detailed discussion.) Rather we want to argue that the teleological account provides an illuminating perspective on two vexed (and we suggest connected) issues in current "theory of mind" research:

1. What is involved in grasping the concept of belief, or (to put the same point differently) in having an "explicit understanding" of belief?
2. What explains the striking dissociations that have been found between children's performance on direct and indirect tests for false belief understanding?

In the following two sections we sketch an answer to (1). Drawing on this account, in the final section, we will tackle (2).

The concept of belief: reasons vs. laws

We can distinguish two aspects of the causal role of beliefs. One has to do with the input side, the circumstances in which beliefs are acquired and sustained. The other concerns the ways beliefs affect what people do. Someone who has acquired a rudimentary "theory of mind" including the concept of belief—who knows what it is to believe something and who is thus able to have thoughts, beliefs, desires, etc., *about* beliefs—must presumably be familiar, to some extent, with both aspects of the causal role of beliefs. We can put the point by formulating "Introduction" and "Elimination" rules for the concept of belief, comparable (in some ways) to the Introduction and

² The adult version of the principle of charity is usually taken to demand that we interpret others in such a way as to make them come out as rational as possible (consistent with the available evidence, of course). See Schueler (2003), Chapter 4, for illuminating discussion.

Elimination rules for the logical constants.³ The claim would be that mastery of the concept of belief requires being able to reason in accordance with these rules. It would be a difficult task to produce a complete list of the relevant rules, but here an example will suffice:

Introduction rule for Belief

A subject S intentionally puts an object O in location L, is not present when O is subsequently moved elsewhere, and has no reason to think O has been moved.

Therefore, S probably believes that O is still at L.

Elimination rule for Belief

S believes O is at L, and decides to retrieve O.

Therefore S will probably make his/her way to L.

What does it come to, being disposed to follow rules such as these? One influential suggestion is that the thinker must have assimilated a psychological theory, consisting of (more or less platitudinous-sounding) law-like generalizations. On this view, there is a sense in which our disposition to reason in accordance with the Introduction and Eliminations Rules for Belief is underpinned by our possession of a simple theory of belief. For example, our use of the Introduction rule reflects our knowledge that if someone puts an object in a certain place, and does not witness its removal from that place, they tend to believe that the object remains in that place.

We can bring out a basic problem with this account by comparing and contrasting the concept of belief with other psychological concepts. Consider the concept of being drunk. Someone who has acquired the concept of drunkenness is someone who is disposed to draw inferences such as the following:

Elimination rule for Drunkenness

Subject S is drunk.

Therefore, S is probably unsteady on his/her feet.

A salient difference between the two cases is this. Why does drunkenness give rise to its familiar symptoms? Why, for example, does it *impair* rather than *boost* our motor skills? To most of us, the matter is deeply opaque. We have no idea *why* the Elimination rule for Drunkenness holds. No doubt there is a story to be told, tracing the effects of alcohol on the motor system. But you don't need to know that story to know what it is to be drunk: the concept of drunkenness is, in that sense, a relatively shallow concept. The concept of belief differs in this respect. It's not only the experts who understand why the Introduction and Elimination rules for Belief hold. Why should S's belief that O is at L induce him to go to L, rather than to dance a jig? The matter is transparent to any reflective thinker who can be said to have a belief-desire psychology (theory of mind): given S's belief and his other circumstances, it *makes sense* for him to go to L—that's where he ought to go (in the "ought of rationality" sense). We have a deeper understanding, in the belief case, of what might be called the rationale of the Introduction and Elimination rules, i.e. the reason why they hold.

³ To illustrate, the Introduction rule for conjunction is

p

q

p & q

For discussion of the relation between understanding and the disposition to reason in accordance with Introduction and Elimination rules (in a range of cases), see Campbell (2002).

Our understanding consists in (a) our ability to reason that S's circumstances (as specified in the Introduction rule) render his belief rational, and that his belief in turn helps to render his action (as specified in the Elimination rule) rational, and (b) our conception of people as rational thinkers and agents. (b) is of course a less austere view than young children's conception of people as responsive to *reasons* (i.e. reason-giving facts). The adult view allows that people may act rationally on the basis of false beliefs and flawed values. And of course we recognize that sometimes people act irrationally.

Belief as perspective: supposition vs. simulation

What's the nature of the reasoning involved in (a)? It is natural, at this point, to turn to the simulation theory. Normally developing humans have a capacity for "imaginative identification" with others. It is in virtue of that capacity, it might be said, that we understand the rationale of the Introduction and Elimination rules. For example, we put ourselves in Mistaken Max's situation, imagine *deciding* to recover the chocolate and *believing* the chocolate to be in the drawer, and then, still within the context of the imaginative exercise, reason to the conclusion "I should go to the drawer." This would be congenial to Jane Heal's and Robert Gordon's views of the role of simulation, which are motivated in part by a concern with the role of rationality in psychological explanation (Gordon, 1995; Heal, 1995). But do we really need to *imagine* having S's mental states to work out that S should go to L? There is a familiar distinction, in the literature on imagination, between supposing and imagining (see, for example, Gendler, 2000; Moran, 1994; Soteriou, 2010). We want to suggest that it is supposition, rather than simulation that holds the key to understanding the rational-explanatory role of beliefs. It's not just that supposition is a more economical procedure than simulation. The important point is that it takes *real* (hypothetical or counterfactual) reasoning, not just imagined or simulated reasoning, to understand what it is rational to do, given the agent's beliefs.

To see the rationale of our Elimination rule for belief it's essential to appreciate that *if* the chocolate were still in the drawer, then this would give Mistaken Max a reason to go to the drawer (i.e. then Max ought—in the "ought of reason" sense—to go to the drawer). To believe that p, after all, is to take it to be a fact that p. And what believing that p makes it rational for one to do depends on what the fact that p would give one a reason to do. Understanding what S's belief makes it rational for him to do thus requires understanding S's *perspective* on what he has reason to do. On the face of it, though, this kind of "perspective taking" is a fairly basic phenomenon—it's not clear that mental simulation necessarily comes into it. The natural way to reach our critical conditional is to reason as follows. "Suppose that the chocolate is still in the drawer. Then what should Max, who urgently needs his chocolate, do? Why, the best course of action, surely, would be for him to go to the drawer." Of course, in one sense, to suppose that p just is to imagine that p. But this is to be distinguished from the richer sense, or senses, of imagination commonly associated with "simulation," such as imagining "from the inside" Max's experiences (e.g. imagining craving chocolate) or the kind of internal play-acting that may be involved in imagining Max's thoughts or propositional attitudes.

Suppositional reasoning involves using as premises propositions one does not believe to be true. But there is nevertheless a sense in which suppositional reasoning is essentially *truth-directed* reasoning. For one thing, we reason from a supposition using the same rules of inference that govern our reasoning from premises we accept. For another, suppositions can be discharged. If you suppose that p, and derive the conclusion that q, you won't of course accept outright that q; but you will, or should, accept outright that *if* p then q. As Dummett puts it: "the point of the procedure [is] that from the fact that certain consequences follow from some hypothesis, we can draw a

conclusion that no longer depends on that hypothesis” (Dummett, 1981, 309). In our example, the *consequence* that follows from the supposition that the chocolate is still in the drawer is that Mistaken Max has a reason to go to the drawer/ought to go to the drawer (in the “ought of reason” sense). The *conclusion* one draws from this is that *if* the chocolate were still in the drawer, Max would have a reason to go to the drawer. This conclusion, in turn, can be used to establish what it is rational for Mistaken Max to do: given that he *believes* the content of our supposition, he ought to go to the drawer (in the “ought of rationality” sense).

One way in which suppositional reasoning differs from simulation is that it is *third-personal*. To determine what someone else would have reason to do under certain suppositions it is not necessary to “recreate” or “replicate” the agent’s first personal deliberation. One may think of the agent from a third- (or second-) person perspective: what would he (or you) have reason to do under those suppositions? In contrast, simulating practical reasoning, as standardly conceived, is a matter of imagined, or “make believe,” reflection on the question “what should *I* do?” (see Gordon, 1996, 62).

There are a number of considerations to suggest that rational explanation requires suppositional reasoning, rather than imaginative identification. First of all, the idea that it takes an imaginative re-enactment of Max’s thought processes to pass a humble false belief task seems suspect on phenomenological grounds. It’s not clear, furthermore, why imagination, in the rich sense, should be needed to work out where it makes sense for Max to go, given his belief: straightforward suppositional reasoning seems perfectly adequate to that task.⁴ Most importantly, such reasoning would seem to be essential even if, in addition, one performs a practical simulation of Max’s deliberation. For imagining someone’s reasoning to the conclusion that he should do *x* does not commit one to the view that it makes sense for him to do *x*: one can imaginatively re-enact reasoning one takes to be not just based on false premises, but to be confused or deranged. Insofar as a practical simulation is to enable one to appreciate the rationality of Max’s action, it has to reflect one’s independent judgment that, given Max’s belief, it’s rational for him to go to the drawer. That judgment cannot itself be based on simulation. It requires reasoning to the conclusion that, if the chocolate were in the drawer, there would be a justifying reason in favor of Max’s going there; and that therefore, given his belief that the chocolate *is* in the drawer, he ought to go there (in the “ought of rationality” sense). There are two features of such reasoning that bear emphasis. One is that it embeds the simple kind of teleological reasoning at which (we argued) even young children are quite proficient. The other feature is that it requires the reasoner to reason (to *genuinely* reason—not just to *pretend* reasoning) from premises she regards as false, in order to derive true conclusions concerning what’s rational for others to do. We call such reasoning “teleology-in-perspective,” to highlight both its continuity with young children’s simple teleology and the fact that it presents its practitioners with a *perspective problem*. They need to be able to move back and forth between two conflicting points of view on what someone has reason to do.

To sum up our discussion so far: young children’s performance on classical false belief tests reflects both a vital insight—people generally do what it makes sense for them to do—and a crucial limitation—their inability to understand that it can be rational for someone to do something even if there is no objective reason for them to do it. This limitation disables young children from fully grasping the concept of belief: they are unable to understand why believing something has the causal role it does, i.e. to recognize what we called the “rationale” for the Introduction and Elimination rules. The next question is this: how might this account help to shed light on the dissociation between children’s performance on direct and indirect tests?

⁴ See Millar (2004) for illuminating discussion of this point.

Understanding the dissociation: theory vs. teleology

Children's understanding of the role of belief in intentional action has been intensively investigated with the "Mistaken Max" false belief task (Wimmer & Perner, 1983). When Max returns looking for his chocolate, 3-year-old children answer with the actual location (cupboard), while 5-year-olds answer with the location Max believes the chocolate to be in (drawer). Many studies tried to find ways of demonstrating earlier understanding, but a large meta-analysis (Wellman, Cross, & Watson, 2001) of these studies showed that the understanding that action depends on belief develops around 4 years of age.

A dissociation

Clements & Perner (1994) found a dissociation between different measures of understanding. The paradigm was slightly changed. Sam the Mouse used different exits from his abode when looking in one or the other of two boxes outside. He had put a piece of cheese in one box (box 1), then went inside to sleep. While asleep someone transferred his cheese to the other box (box 2). This set up allowed the filming of the children's eye gaze when Sam woke up with a craving for his cheese. Most 3-year-olds looked for Sam in expectation of his reappearance at the exit to box 1 (where he thought his cheese was). This occurred only in the false belief condition, but not in a true-belief control condition, where Sam had seen the transfer to box 2. Most interestingly, all the young 3-year-olds who showed this looking behavior still maintained, when asked, that Sam would come out from the exit to box 2 (where the cheese actually was).

This dissociation has been replicated (Garnham & Perner, 2001) by different investigators (Low 2010; Ruffman, Garnham, Import, & Connolly, 2001; Wang, Low, Jing, & Qinghua, 2012). Perner & Clements 2000 made a case that children's anticipatory looking shows the characteristics of indirect measures indicative of implicit knowledge (Reingold & Merikle 1988), e.g. guessing by blindsight patients of the location of a stimulus in their blind field (Weiskrantz, 1986), by sighted persons in the Roelofs's induced motion illusion (BridgemanKirch, & Sperling, 1981; Bridgeman, Peery, & Anand, 1997), thumb-finger span size indicating an object's true size when explicit size judgments are distorted by illusion effects (Agliotti, DeSouza, & Goodale, 1995; Stöttinger, Aigner, Hanstein, & Perner, 2009; Stöttinger, Soder, Pfusterschmied, Wagner, & Perner, 2010).⁵ Furthermore, Ruffman et al. (2001) showed that children seem absolutely unaware—don't even

⁵ The distinction between direct and indirect tests is not as obvious as it may seem. Naively one would think that a direct false belief test is one in which the child is asked directly about an agent's belief. In that case, the good old standard false belief test would be indirect, because children are not asked about Mistaken Max's belief, but about his future action. Hence, the test should strictly speaking not be called a direct false belief test, but a direct test of mistaken intentional action.

In general other problematic aspects are that the question may be directly about the matter of interest but still count as indirect. For instance, when a blindsight person is asked to guess where a stimulus is, implicit knowledge can be used, but not when asked to point to where the stimulus actually is. The same can be shown with normally sighted persons when they have to indicate a near threshold change of brightness (Marcel, 1993). So the critical feature is not the form of the question but how the question is to be taken. If the respondent is to take it as a request to say where something really is, then it is a direct test. If the question is to be taken as where something could be (a blind guess), then it is an indirect test, because the pointing gesture to where it could be (guess) is influenced by where it actually is. Moreover, when blindsight patients are asked to insert their hand into a slot of different orientation they can do so above chance even when they can't consciously see the slot, but they cannot indicate with their hand the direction of the slot (Perenin & Rosetti, 1996). Similar abilities have been reported with healthy persons' susceptibility to illusions.

consider it a vague possibility—that the agent could reappear where they look in anticipation to see him reappear.

Early sensitivity—the facts

The use of indirect tests has led to the discovery of very early sensitivity to agent's false beliefs. We can distinguish four different paradigms.

Looking in expectation

Children look in expectation where they expect a hand to appear on the basis of where the agent thinks an object is. This can be shown by 2 years (Southgate, Senju, & Csibra, 2007) and perhaps earlier (Neumann, 2009; Southgate, 2008).

Looking time

Infants of about 14 months look longer at the test scene when a mistaken agent searches in the correct location than when she searches in the wrong location where she thinks the object is (Onishi & Baillargeon, 2005). The longer looking is interpreted as infants expecting a different action (search in the empty container where the agent believes the object to be) than what is shown (agent searches in correct container). Hence this method has been dubbed “violation of expectation paradigm.” This finding has led to an explosion of demonstrations that infants in their second year expect agents to act according to their beliefs and not the real state of affairs.

A somewhat different use of looking time differences was made by Kovacs, Teglas, & Endress (2010). As early as 7 months, infants' looking time was recorded when discovering a surprising outcome—a ball behind a screen had disappeared. Their looking time was longer when a bystander shared their belief that the ball was still behind the screen than when the bystander thought the ball had disappeared. A similar technique using reaction times has been pioneered by Apperly, Riggs, Simpson, Chiavarino, & Samson (2006) assessing automaticity of belief attribution in adults with the conclusion that it is not automatic.

Interpretation of referential expressions

This paradigm was pioneered by Carpenter, Call & Tomasello (2002) and Happé & Loth (2002) with children around the age of 3 years. Southgate, Chevallier, & Csibra (2010) tested 17-month-old infants who watched an agent place two novel, unnamed objects in two separate boxes. Unbeknownst to the agent, the contents were then switched. When the agent returned, she pointed to a box (the incorrect box in the false-belief condition) and said: “Do you remember what I put in here? There's a *sefo* in here. There's a *sefo* in this box. Shall we play with the *sefo*?” In the false-belief conditions, children correctly chose the item in the other box, not the one the experimenter pointed to. The authors' interpretation is that children understood that the experimenter wanted from the indicated box the object she thought was in there and not the one that was actually in there.

Helping behavior

In the false-belief condition by Buttelmann, Carpenter, and Tomasello (2009), an agent failed to witness her favourite toy being moved and returned to the first box to retrieve the toy, but couldn't get to open the box. Children were then asked to “help” the agent. Over 70% of 18-month-olds approached the second box. In contrast, less than 20% did so in a knowledge condition where the unsuccessful agent had witnessed the transfer to the new box but tried to open the empty box. Buttelmann et al. (2009) suggested that toddlers approached the second box in the false-belief condition because they recognized that the agent falsely believed that her toy was still inside the first box and concluded from the agent's unsuccessful attempt to open that box that she wanted to

retrieve the toy she thought was in that box but which was now in the new box. So the child had to orient to the new box to retrieve the desired toy.

Early sensitivity—interpretation

One question about these findings concerns the best way to characterize the two groups of tasks—those that reveal early sensitivity and those traditional tasks that point to later understanding. Clements and Perner (1994; Perner & Clements, 2000) characterized the tasks as indirect and direct, inspired by the use of this terminology in the consciousness literature (Reingold & Merikle, 1988). More recently, Scott and Baillargeon (2009) characterized the difference as one of “spontaneous” and “elicited” responses, which has much to recommend itself, but also does not quite capture the relevant difference, as the authors themselves imply (p. 391): “Finally, infants and toddlers should succeed at indirect-elicited-response tasks that require them to respond to questions or prompts that only indirectly tap their representation of an agent’s false belief.” And the authors refer to the studies by Buttelmann et al (2009) and Southgate et al (2010) as good examples. So “indirectness” seems the critical factor.

Several distinctions have been proposed to characterize the difference between the kinds of knowledge underlying the early sensitivity and later understanding:

1. Implicit—explicit (Clements & Perner, 1994; Perner & Clements, 2000)
 - (a) unconscious—conscious (Garnham & Perner, 2001; Ruffman et al., 2001)
 - (b) procedural—declarative
 - (c) non-conceptual—conceptual (Rakoczy, 2012)
 - (d) automatic (spontaneous)—controlled (Apperly et al., 2006)
2. Modular—central process (Leslie, 1994)
3. Causal understanding: shallow—deep (behavior rules—mental state rules; Perner, 2010)

But from these characterizations no general principles follow for a detailed account, which would answer the following two questions:

1. Why does early sensitivity emerge only in indirect tasks and not the traditional direct ones?
2. Why do the younger children give systematically wrong answers on the direct tests until they are about 4 years old?

One detailed account was provided by Scott and Baillargeon (2009). They propose two subsystems to the theory of mind system SS1 and SS2 (pp. 1174–5):

SS1 allows them to attribute two kinds of internal states to the agent: motivational and reality-congruent informational states ... Motivational states specify the agent’s motivation in the scene and include goals and dispositions. Reality-congruent informational states specify what knowledge or accurate information as construed by the infant the agent possesses about the scene... SS2 extends SS1 in that it allows infants to attribute reality-incongruent informational states to agents. When an agent holds a false or a pretend belief about a scene ...

Scott and Baillargeon’s answer to our two questions rests on the assumption that indirect tests require only representation of belief, while direct tests require the interplay of three processes (p. 1176):

We assume that success in the Sally–Ann task depends on the interaction of three separate processes. First, children must represent Sally’s false belief about the marble’s location; ... Second, when asked the

test question, children must attend to the question, decide to answer it, and tap their representation of Sally's false belief (*response-selection process*). Finally, children must inhibit any prepotent tendency to answer the question based on their own knowledge of the marble's current location (*response-inhibition process*). ... Children then fail because (a) the joint activation of the false-belief-representation process and the response-selection process overwhelms their limited information-processing resources, and/or (b) the neural connections between the brain regions that serve these two processes are still immature and inefficient in early childhood.

This answers question 1: children show sensitivity to false belief in indirect tests earlier than in direct tests because indirect tests tax their limited processing system less than direct tests. An answer to question 2 is not obvious. If the overload leaves the toddler without any means to answer then the child can but guess, but not be systematically wrong. If the overload disables SS2, but leaves SS1 as default, the consequence would be that SS1 would represent the agent's ignorance of the new location and the child, again, can but guess what the agent will do.

We can see two interesting weaknesses in this approach. The one, already discussed, is the need to explain the systematic errors on the direct tests. The other weakness is that it does not explain why only direct tests require response selection and inhibition but not also indirect tests. Scott, Baillargeon, Song, & Leslie (2010, p. 391) have this to say:

In marked contrast, success in spontaneous-response tasks such as VOE [violation of expectation] and AL [anticipatory looking] tasks depends on only one process, the false-belief-representation process; the response-selection and response-inhibition processes are not activated because children produce their responses spontaneously rather than in answer to direct questions.

So the response is given spontaneously without external prompt by a question,⁶ but that still leaves the child to select one of many responses (e.g. should I look to location 1 or to location 2 if I want to see him come out of the exit). Selection of the correct looking response would also be interfered with by the tendency to look to the exit near the desired object's real location, which needs to be inhibited. In fact, if selection of the spontaneous looking response tends to be automatic and implicit in contrast to an explicit response to a question, then it should, if anything, be more difficult to inhibit the automatic response than the more explicit response—exactly the opposite of what is being observed.

From causes of behavior to reasons for acting as causes

Our proposal also assumes two different approaches (or systems). One consists of a purely nomic causal understanding of *behavior* as caused by motivational and informational states in the tradition of theory theory (Gopnik and Meltzoff, 1997; Leslie, 1994). It underlies the data based on indirect tests. The other consists of understanding *reasons for action* in the tradition of those who emphasize the role of rationality (and the principle of charity) in interpretation (Davidson, 1963; McDowell, 1985). It is triggered by direct tests and is based on expecting people to act in a way they have reasons to act.

Approach 1: caused behavior

For present purposes we'd like to remain completely agnostic about the causal depth of this understanding, e.g. whether infants make causally shallow connections from observable indicators of

⁶ In fact the prompt need not be a question. Garnham and Perner (2001) had children place a mat to catch the returning agent. Children who placed the mat spontaneously without hesitation tended to place the mat to the exit where the agent thought his object was, while children who needed prompting in the form of a reminder to move the mat tended to place it at the exit where the object really was.

motivational and informational states to future behavior (behavior rules, Povinelli & Vonk, 2004) or whether they infer inner states from these indicators, and predict and interpret behavior on the basis of these inner states (e.g. Tomasello, Call, & Hare, 2003).⁷ We also stay neutral about origin. Infants' knowledge might be innate (and modular) and emerge at particular times by maturation (Leslie, 1994), or be rapidly built up by statistical learning (Ruffman, Taumoepeau, & Perkins, 2012) or be acquired through a theorizing process (Gopnik & Meltzoff, 1997).

We have commitment—though not irrevocable—on some of the other features of its knowledge base. Evidence suggests that it is based on implicit knowledge. The dissociation observed by Clements and Perner (1994) and children's reluctance to acknowledge the agent to reappear where they looked in anticipation (Ruffman et al., 2001) suggests that it is not consciously accessible. Consequently, it is likely to be automatic and not under voluntary control (Apperly & Butterfill, 2009). Moreover, indirect measures tend to consist of online "reactions to the unfolding events" (Scott and Baillargeon, 2009, p. 1176) or live interactions in the helping paradigms. This also suggests that the knowledge is procedural and may not be available for conditional reflection, which requires declarative knowledge.

The general characterization of this approach is that it treats behavior of organisms or moving dots on a par with the movement and changes of inanimate physical objects. Theory of mind is just one theory among many others.

Approach 2: reasons for action as causes

To appreciate behavior (goal-directed movement) as intentional action, one has to understand it as behavior for which the agent has reasons. If one were in Mistaken Max's situation, having put one's chocolate earlier into the drawer and now looking for it, one does not ask oneself what one *will* do next, but what one *should* do next—go to the drawer or the cupboard? Since from one's own point of view, the chocolate is still in the drawer, one seems to have good (objective) reasons to go to the drawer. And because one is motivated by this fact, one knows that one is likely to go there because one should go there. One does not simply conclude this on the basis of a law-like regularity: "Whenever I want something and think (know) that it is in location x then I will go to location x."

To understand others as intentional agents is to understand what they are doing or will do in terms of what they should do given their goals and circumstances.⁸ Rakoczy, Warneken, and Tomasello (2008) showed that children as young as 2 years expect other people to act as they should. If the pronounced goal is to play a certain game then one should behave according to the game's constituent rules. Or else 2- and particularly 3-year-olds get very upset. Children's normative attitude is based on understanding objective reasons, i.e. teleology (Perner & Roessler, 2010). Teleology captures intentional actions very well, as long as they are based on objective goals and objectively appropriate instrumental actions.⁹

⁷ Despite the recent evidence of early understanding of the mind the critical evidence whether this understanding is based on behavior rules or mental state computation is still outstanding (Perner, 2010).

⁸ We want to emphasize (see above Belief as perspective: supposition vs. simulation) that this should not be understood in the sense of simulation theory as imaginatively putting oneself into the other person's situation. It only requires seeing from one's own position what is needed (goal) and what needs to be done to achieve it (instrumental action) by whoever is in a position to carry out that action.

⁹ This bears resemblance to Scott and Baillargeon's (2009) subsystem SS1 as it involves goals and—to use their term—reality-congruent instrumental actions.

Teleology breaks down as a means of understanding intentional actions when subjective perspectives on goals and means are involved. Mistaken Max's move to the empty drawer remains an irrational behavior for the young teleologist. The teleologist can, however, recapture Mistaken Max's rationality by realizing that he is mistaken, i.e. has a deviant perspective on the world, and employ teleology within his perspective (Perner, 2004). Earlier we labelled this kind of reasoning "teleology in perspective" (see Belief as perspective: supposition vs. simulation; see also Perner & Roessler, 2010). Importantly, teleology-in-perspective preserves the rationality of Mistaken Max's action: he can be seen to act on the basis of what from his perspective appears to be an objective reason. This ability becomes operative around 4 years when children develop some notion that different perspectives exist and, thus, need not anymore rely on being switched to another person's perspective but can voluntarily do so (Perner, Stummer, Sprung, & Doherty, 2002).

There is now a large amount of evidence that children at this age become able to succeed on a large variety of otherwise unrelated tasks that share the need for perspective understanding. For instance, level 2 perspective taking (Masangkay, McCluskey, McIntyre, Sims-Knight, Vaughn, & Flavell, 1974), interpreting ambiguous drawings (Doherty & Wimmer, 2005), understanding false direction signs (Leekam, Perner, Healey, & Sewell, 2008) alternative naming (Doherty & Perner, 1998; Parkin, 1994; Perner et al., 2002), episodic memory (Perner & Ruffman, 1995; Perner, Kloos, & Stöttinger, 2007; Sabbagh, Moses, & Shiverick, 2006), and understanding identity information (Perner et al., 2010) not only emerge at this age, but also correlate specifically with the traditional false belief task.

A main purpose of understanding reasons for action is to explain, rather than predict behavior (Andrews, 2012) and be able to reason and argue about the correctness or appropriateness of one's own and others' conduct. It is an essential glue of human society and tied to linguistic interaction. Its knowledge base must be explicit: declarative (non-procedural) to be used in conditional arguments, access conscious, conceptual (for linguistic exchanges), and under voluntary control (at least for voluntary retrieval). It cannot be modular, since it needs to be accessible for argumentation.

Explaining the evidence: answering our two questions

Having described the two approaches taken by children (also by adults) we need to check how well this proposal can answer our two questions (from the section Early sensitivity—interpretation):

1. Children show sensitivity to false beliefs very early on indirect measures in online and interactive tasks because they have implicit nomic knowledge about motivational and informational states causing behavior (we are non-committal as to the causal depth of this knowledge). In contrast, on direct tests the knowledge in question is part of the test specification ("Where will Max go?") which requires a declarative commitment, which—as the consciousness literature suggests—requires *explicit knowledge*. For this the young children employ pure teleology; they make predictions of what someone will do in terms of what the person should do, i.e. has objective reasons to do.
2. Children's predictions on direct false belief tests show the reality error because they are teleologists. Around 4 years they become aware of the existence of perspective differences, which enables them to see a person's reasons relative to a different perspective (teleology in perspective). The age point conforms to the age at which many other tasks that require awareness of perspectives are mastered.

Conclusion

We drew attention to a feature of how we understand intentional action that tends to get lost in the theory theory of mind. Our naïve belief-desire psychology is not primarily a body of law-like generalizations of how agents tend to behave, but involves an understanding that they act for reasons. They do what they should do—for the most part. We ventured the contention that infants' and toddlers' early expectations of how people will act, especially when a false belief is involved, may be based on law-like generalizations, which remain implicit and dissociate from an understanding of people acting for reasons until they become able to understand reasons relative to an agent's perspective around 4 years. Because of the dissociation, we think that the earlier understanding is implicit and the later understanding explicit. Our mentalist understanding (theory of mind) of intentional actions has to become explicit at some point anyway, since one of its prime functions is to argue about and justify conduct. Although we emphasized the difference in knowledge base of earlier and later understanding, we also like to think that there is developmental continuity between the approaches (this is one reason we do not want to talk of systems that are often associated as independent); there is evidence that performance in direct tasks is related to earlier performance on indirect tasks (Low, 2010; Thoermer, Sodian, Vuori, Perst, & Kristen, 2011). In particular, the discrepancy between how mistaken people act and the young teleologist's wayward predictions must be an important motor for moving to a more sophisticated understanding. As Karmiloff-Smith and Inhelder (1974) observed in the context of children's understanding of how to balance objects on a fulcrum: "If you want to get ahead, get a theory." So, in our case, infants have an implicit sense of how people under certain informational conditions are likely to act, then they get a rough theory (teleology) that people act as they should act, which they then need to refine into teleology in perspective.

References

- Agliotti, S., DeSouza, J. F., & Goodale, M. A. (1995). Size-contrast illusions deceive the eye but not the hand. *Current Biology*, 5(6), 679–85.
- Andrews, K. (2012). *Do Apes Read Minds: Toward a New Folk Psychology*. Cambridge: MIT.
- Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116, 953–70.
- Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, 17, 841–4.
- Bridgeman, B., Kirch, M., & Sperling, A. (1981). Segregation of cognitive and motor aspects of visual function using induced motion. *Perception and Psychophysics*, 29, 336–42.
- Bridgeman, B., Peery, S., & Anand, S. (1997). Interaction of cognitive and sensorimotor maps of visual space. *Perception and Psychophysics*, 59(3), 456–69.
- Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112(2), 337–42.
- Campbell, J. (2002). *References and Consciousness*. Oxford: Oxford University Press.
- Carpenter, M., Call, J., & Tomasello, M. (2002). A new false belief test for 36-month-olds. *British Journal of Developmental Psychology*, 20, 393–420.
- Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, 9, 377–97.
- Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, 60, 685–700.
- Doherty, M. J., & Perner, J. (1998). Metalinguistic awareness and theory of mind: Just two words for the same thing? *Cognitive Development*, 13, 279–305.

- Doherty, M. J., & Wimmer, M. (2005). Children's understanding of ambiguous figures: Which cognitive developments are necessary to experience reversal? *Cognitive Development*, 20, 407–21.
- Dummett, M. (1981). *Frege. Philosophy of Language*, 2nd edn. London: Duckworth.
- Garnham, W. A., & Perner, J. (2001). When actions really do speak louder than words—but only implicitly: Young children's understanding of false belief in action. *British Journal of Developmental Psychology*, 19, 413–32.
- Gendler, T. S. (2000). The puzzle of imaginative resistance. *Journal of Philosophy*, 97, 55–80.
- Gopnik, A., & Meltzoff, A. N. (1997). *Word, Thoughts, and Theories*. Cambridge: Bradford Books, MIT Press.
- Gordon, R. (1995). Folk psychology as simulation. In M. Davies, & T. Stone (Eds), *Folk Psychology*. Oxford: Blackwell.
- Gordon, R. M. (1996). “Radical” simulationism. In P. Carruthers, & P. K. Smith (Eds), *Theories of Theories of Mind*, pp. 11–21. Cambridge: Cambridge University Press.
- Happé, F., & Loth, E. (2002). “Theory of mind” and tracking speakers’ intentions. *Mind and Language*, 17, 24–36.
- Heal, J. (1995). “Replication and Functionalism”. In M. Davies & T. Stone (Eds), *Folk Psychology*. Oxford: Blackwell.
- Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition*, 3, 195–212.
- Kolodny, N. (2005). Why be rational? *Mind*, 114, 509–63.
- Kovacs, A. M., Teglas, E., & Endress, A. D. (2010). The social sense: susceptibility to others’ beliefs in human infants and adults. *Science New York, NY*, 330(6012), 1830–4.
- Leekam, S., Perner, J., Healey, L., & Sewell, C. (2008). False signs and the non-specificity of theory of mind: Evidence that preschoolers have general difficulties in understanding representations. *British Journal of Developmental Psychology*, 26(4), 485–97.
- Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition*, 50, 211–38.
- Low, J. (2010). Preschoolers’ implicit and explicit false-belief understanding: relations with complex syntactical mastery. *Child Development*, 81(2), 597–615.
- Marcel, A. J. (1993). Slippage in the unity of consciousness. In G. R. Bock, & J. Marsh (Eds), *Experimental and Theoretical Studies of Consciousness* (pp. 168–86). Chichester: Wiley.
- Masangkay, Z. S., McCluskey, K. A., McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The early development of inferences about the visual percepts of others. *Child Development*, 45, 357–66.
- McDowell, J. (1985). Functionalism and anomalous monism. In E. LePore, & B. McLaughlin (Eds), *Actions and Events*. Oxford: Blackwell.
- Millar, A. (2004). *Understanding People*. Oxford: Clarendon Press.
- Moran, R. (1994). The expression of feeling in imagination. *Philosophical Review*, 103, 75–106.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308, 255–8.
- Neumann, A. (2009). Infants’ implicit and explicit knowledge of mental states. *Evidence from Eye-tracking Studies*, Chapter 7. Ketsch Verlag: Microfiche.
- Parkin, L. J. (1994). *Children's understanding of misrepresentation*. Unpublished Thesis for doctorate, University of Sussex.
- Perenin, M. T., & Rossetti, Y. (1996). Grasping without form discrimination in an hemianopic field. *Neuroreport*, 7, 793–7.
- Perner, J. (2004). Wann verstehen Kinder Handlungen als rational? In H. Schmidinger & C. Sedmak (Eds), *Der Mensch—ein“animal rationale”? Vernunft—Kognition—Intelligenz* (pp. 198–215). Darmstadt: Wissenschaftliche Buchgesellschaft.

- Perner, J. (2010). Who took the cog out of cognitive science? Mentalism in an era of anti-cognitivism. In P. A. Frensch, & R. Schwarzer (Eds), *Cognition and Neuropsychology: International Perspectives on Psychological Science*, Volume 1 (pp. 241–61). London: Psychology Press.
- Perner, J., & Clements, W. A. (2000). From an implicit to an explicit theory of mind. In Y. Rossetti, & A. Revonsuo (Eds), *Beyond Dissociations: Interaction Between Dissociated Implicit and Explicit Processing* (pp. 273–93). Amsterdam: John Benjamins.
- Perner, J., Kloos, D., & Stöttinger, E. (2007). Introspection and remembering. *Synthese*, 159, 253–70.
- Perner, J., Mauer, M. C., & Hildenbrand, M. (2011). Identity: Key to children's understanding of belief. *Science*, 333, 474–7.
- Perner, J., & Roessler, J. (2010). Teleology and causal reasoning in children's theory of mind. In J. Aguilar, & A. Buckareff (Eds), *Causing Human Action: New perspectives on the causal theory of action* (pp. 199–228). Cambridge: Bradford Books, MIT Press.
- Perner, J., & Ruffman, T. (1995). Episodic memory an autoeic consciousness: Developmental evidence and a theory of childhood amnesia. Special Issue: Early memory. *Journal of Experimental Child Psychology*, 59(3), 516–48.
- Perner, J., Stummer, S., Sprung, M., & Doherty, M. (2002). Theory of mind finds its Piagetian perspective: Why alternative naming comes with understanding belief. Cognitive Development, Special Issue on "Constructivism Today" as part of the inauguration of Cognitive Development as the official journal of the Jean Piaget Society for the Study of Knowledge and Development. *Cognitive Development*, 17, 1451–72.
- Povinelli, D. J., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind & Language*, 19, 1–28.
- Rakoczy, H. (2012). Do infants have a theory of mind? *British Journal of Developmental Psychology*, 30, 59–74.
- Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: young children's awareness of the normative structure of games. *Developmental Psychology*, 44, 875–81.
- Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics*, 44(6), 563–75.
- Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology*, 80, 201–24.
- Ruffman, T., Taumoepeau, M., & Perkins, C. (2012). Statistical learning as a basis for social understanding in children. *British Journal of Developmental Psychology*, 30, 87–104.
- Sabbagh, M. A., Moses, L. J., & Shiverick, S. (2006). Executive functioning and preschoolers' understanding of false beliefs, false photographs and false signs. *Child Development*, 77, 1034–49.
- Schueler, F. (2003). *Reasons and Purposes*. Oxford: Clarendon.
- Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development*, 80(4), 1172–96.
- Scott, R. M., Baillargeon, R., Song, H. J., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology*, 61(4), 366–95.
- Soteriou, M. (2010). Cartesian reflections on the autonomy of the mental. In J. Aguilar, A. Buckareff & K. Frankish (Eds), *New waves in philosophy of action* (pp. 122–40). Palgrave: Macmillan.
- Southgate, V. (2008). "Attributions of false belief in infancy." Invited presentation at the EPS Research Workshop on Theory of Mind: A workshop in celebration of the 30th anniversary of Premack & Woodruff's seminal paper, "Does the chimpanzee have a Theory of Mind? (BBS 1978), University of Nottingham, 11–12 September
- Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, 13(6), 907–12.
- Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18, 586–92.

- Stöttinger, E., Aigner, S., Hanstein, K., & Perner, J. (2009). Grasping the diagonal: controlling attention to illusory stimuli for action and perception. *Consciousness and Cognition*, 18(1), 223–8.
- Stöttinger, E., Soder, K., Pfusterschmied, J., Wagner, H., & Perner, J. (2010). Division of labour within the visual system—fact or fiction?—Which kind of evidence is appropriate to clarify this debate? *Experimental Brain Research*, 202, 79–88.
- Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2011). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology*, 30, 172–87.
- Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees versus humans: It's not that simple. *Trends in Cognitive Sciences*, 7, 239–40.
- Wang, B., Low, J., Jing, Z., & Qinghua, Q. (2012). Chinese preschoolers' implicit and explicit false-belief understanding. *British Journal of Developmental Psychology*, 30, 123–140.
- Weiskrantz, L. (1986). *Blindsight: A Case Study and Implications*. Paperback edn 1998. Oxford: Oxford University Press.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development*, 72, 655–84.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–28.
- Woodward, J. (2011). Causal perception and causal cognition. In J. Roessler, H. Lerman & N. Eilan (Eds), *Perception, Causation, and Objectivity* (pp. 229–63). Oxford: Oxford University Press.

Chapter 4

Theory of mind, development, and deafness

Henry M. Wellman and Candida C. Peterson

Philosophers and psychologists often characterize our everyday system of reasoning about mind, world, and behavior as a belief-desire psychology (D'Andrade, 1987; Fodor, 1987; Wellman, 1990). Such an everyday psychology, often termed a theory of mind, provides explanations and predictions of intentional action by appeal to what the person thinks, knows, and expects coupled with what he or she wants, intends and hopes for. Why did Jill go to the drawer: She *wanted* her chocolate and *thought* it was in the drawer. Everyday psychology also includes reasoning about the origins of mental states (Jill wants candy because she is *hungry*; Jill thinks it is in the drawer where she last *saw* it). That is, everyday or naïve psychology incorporates a variety of related constructs, such as drives and preferences that ground one's desires, and perceptual-historical experiences that ground one's beliefs. It includes emotional reactions that result from these desires, beliefs, intentions, actions, and perceptions—happiness at fulfilled desires, frustration at unfulfilled desires, surprise when events contradict one's firmly held beliefs. We consider how children develop such a theory of mind.

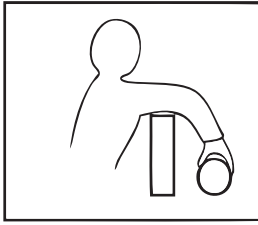
For background, and because it is important in its own right, we begin with a brief overview of theory of mind development in typically developing children. As follows from the above outline of everyday psychological reasoning, children's developing understanding of beliefs, desires, perceptions, intentions, and emotions, are all of interest and importance (see Harris, 2006). Our overview sets the stage for two foci we consider in more depth. The first concerns the use of new, insightful methods to more deeply examine how theory of mind unfolds over development. Because theory of mind is a developmental achievement, we argue that research must more richly and deeply reveal how it actually develops. The second focus concerns the insights to be achieved about theory of mind from atypical development. Here, we argue that, although atypical developments in children with autism have been most extensively studied (as reviewed elsewhere in this volume), developments in children with deafness are both complementary and especially revealing.

The course of theory of mind development

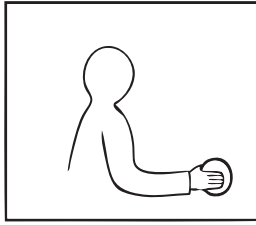
Development of an understanding of people is a lifelong task, beginning at birth. Infants who are only a few days old prefer to look at people and faces, imitate people, but not inanimate devices, listen to human voices, and so on. More pertinently, by the end of the first year children begin to treat themselves and others as intentional agents and experiencers.

Box 4.1 provides an example paradigm used to demonstrate intention understanding in infants (from Brandone & Wellman, 2009; Phillips & Wellman, 2005). In early demonstrations using something like this paradigm, infants saw an animated circle “jumping” over a barrier to reach its goal-object. Just as they do for intentional human acts, 9- and 12-month-olds look longer at

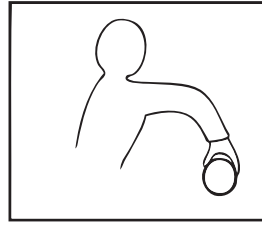
Box 4.1



Habituation event



Direct reach test event



Indirect reach test event

Habituation-test (or familiarization-test) paradigms are designed so that participants will look longer at novel, unexpected test events more than at familiar, expected test events. In the reaching paradigm (depicted above), during habituation, participants view multiple trials of the agent reaching over the barrier for the goal object. Then, the barrier is *removed* and the test events contrast two different construals of the person's actions, one in terms of intentions and one in terms of physical motions of the body. If during habituation participants construe the action in terms of its physical movement (the arcing arm motion), then the indirect reach test event should be expected (as it repeats the same movement) whereas the direct reach will stand out as novel and so especially attention-worthy. In contrast, if participants initially construe the action as goal directed (the actor going as directly as possible to get her goal), then when the barrier is removed the direct reach would be the expected action because the actor continues to directly seek the goal, and the indirect reach would be more attention-worthy because (although the actor's arm movement remains the same as during habituation) the action is no longer straightforwardly directed to the goal. In this paradigm, 8-, 9-, 10-, and 12-month-olds, and chimpanzees and macaques consistently look longer at the *indirect* test event.

the indirect test event, showing an abstract, generalized understanding of goal-directed agency (Csibra, Gergely, Biro, & Brockbank, 1999).

Relatedly, intentional action is not only directed toward specific goals, it is non-accidental (e.g. Carpenter, Aktar & Tomasello, 1998a; Olineck & Poulin-Dubois, 2005). To illustrate, in Carpenter et al. (1998a), 14- and 18-month-old infants watched an adult model do several two-action sequences on complex objects (e.g. pushing a button and then moving a lever). One action was marked vocally by the adult as intentional ("There!"), and one as accidental ("Whoops!"). Infants imitated almost twice as many intentional as accidental actions and only very rarely imitated the entire two-action sequences.

When viewing actions, such as those in Box 4.1, or such as pushing buttons and moving levers, infants might conceivably identify only the spatial-directedness and objective efficiency of the overt behavior toward its overt target—a teleological or behavioral, rather than intentional understanding (Gergely & Csibra, 2003). However, inferring a goal when it is unfulfilled (and thus non-overt in the actor's movements or outcomes) demonstrates an understanding that intentions exist beyond the surface actions performed. So, in a seminal study (Meltzoff, 1995), 18-month-olds witnessed an adult try, but fail to fulfill several novel, object-directed goals (e.g. trying to hang a ring on a hook). Although infants never saw the actions successfully modeled, when given a chance to act on the objects themselves they "imitated" the successful action much more than the failed

(actually witnessed) actions. Fifteen-, but not 12-month-olds also display this pattern (Carpenter, Nagell & Tomasello, 1998b).

Motoric imitation is arguably a demanding response system, more so than attentive looking. So, consider a version of the displays in Box 4.1 where the actor reaches for, but falls short of successfully grasping the target object. If habituated to such unsuccessful actions 10- and 12-month-olds (and perhaps 8-month-olds), interpret the actions in terms of the (never actually seen) intentional goal of grasping the object (Brandone & Wellman, 2009; Hamlin, Hallinan, & Woodward, 2008).

Not only do persons engage in intentional action, they experience the world. Infants can appreciate not only intentional action, but also intentional attention and experiences. Potentially, gaze-following, where infants follow an agent's line of sight toward an object, would be produced by an understanding that the agent sees something—the person has a visual experience. Infant gaze following, however, could be “behavioral,” tracking others' head-eye orientations might simply yield for the infant interesting sights without a recognition of the agent's visual experience (Baldwin & Moses, 1996). Yet, by 12–14 months infants also follow an adult's gaze around a barrier, and do so even when this requires leaning or moving behind the barrier (Butler, Caron & Brooks, 2000; Dunphy-Lelii & Wellman, 2004; Moll & Tomasello, 2004).

However, even in appropriately gazing around barriers, infants could be responding to the agent's eye-ball (or head-nose) orientation without a deeper sense of her intentional experience (Moore & Corkum, 1994). For example, at 12 months, infants often “gaze follow” the head turns of adults who wear blindfolds. However, recent data confirm a deeper understanding of visual experience. In Meltzoff and Brooks (2008), 12-month-olds were given advance experience with blindfolds occluding their own vision. After such experiences they were significantly less likely to “gaze follow” a blindfolded adult, suggesting that their sense of what the adult can see—visually experience—guided infants' actions. Eighteen-month-olds do not often gaze-follow a blindfolded adult—probably because they have come to understand that blindfolds occlude visual experience—but in this same study 18-month-olds were given experience with a special blindfold that looked opaque yet was easily seen through when worn. After experience with that blindfold, 18-month-olds did gaze-follow the head turn of a blindfolded adult. Thus, by 12–18 months, it is infants' sense of the person's visual experience (not just overt eye- or head-directedness) that often controls their gaze following.

Persons not only can have intentional experiences about some here and now event, their experiences can accumulate and update (or fail to update) over time. As one example, Tomasello and Haberl (2003) examined this with 12- and 18-month-old infants who interacted with three objects. Critically, a target adult joined in these interactions for two of the objects, but was absent for the third. After these interactions, the target adult saw all three objects displayed on a tray, and said to the infant, “Wow! That's cool! Can you give it to me?” while gesturing ambiguously in the direction of the objects. Three objects were now familiar for the infant, but one was new (and so “cool”) to the target adult. Infants gave the target adult the object that was new for her. Thus, they tracked the adult's experiences sufficiently to know that (a) her experience was not updated when theirs was (a recognition of the subjectivity of experience) and (b) she was previously unaware of (ignorant of) the third object.

Initial infant insights about intention culminate in their understanding that intentional agents behave according to their desires and emotions, constrained by their perceptual experiences. Tomasello, Carpenter, Call, Behne, & Moll (2005) call this “understanding intentions and attention,” Wellman (2011) calls this a “desire-emotion-perception understanding of persons.” Regardless, this infant understanding of persons encompasses a rudimentary, but impressive sense of agents' awareness or unawareness (knowledge or ignorance) of events, a recognition that if

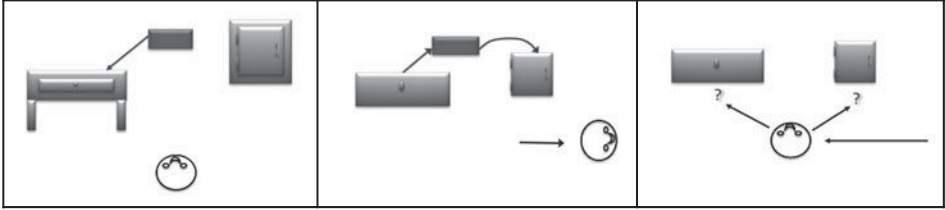
persons' experiences of situations are not updated as events change, then they can be unaware of key circumstances (and thus act in ignorance).

To be unaware of (ignorant about) something is *not* the same as to have a false belief about it. As depicted in Box 4.2, an agent might also have a false belief about (for example) where an object is beyond just being ignorant of its location. False beliefs, when contents of the world (object-in-cupboard) are seen to contradict contents of thought ("object-in-drawer"), provide a powerful, yet everyday, illustration of a "representational" theory of mind. Accordingly, there has been much research on children's understanding of false belief (hundreds of studies in meta-analyses by Milligan, Astington, & Dack, 2007; Wellman, Cross, & Watson, 2001, Liu, Wellman, Tardif, & Sabbagh, 2008). Another reason for this voluminous research is that when researchers were first becoming interested in theory of mind, several easy-to-use "standard" false belief tasks were developed (Box 4.2) and these have proved nicely revealing. Indeed, as shown in Box 4.2, they consistently show an important developmental transition where typically developing children come to an explicit understanding of false belief during the preschool years. This development is representative of broader changes in children's theory of mind. Moreover, because they have been used worldwide, false belief tasks begin to reveal a universal childhood theory-of-mind achievement, again as shown in Box 4.2.

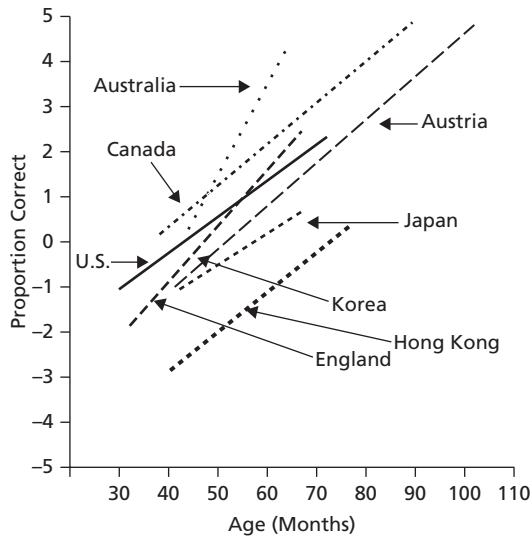
This picture of false belief development as concentrated in the preschool years represented the consensus view, until about 5 years ago. Since then, however, there have been emerging findings claiming that even 12-, 15-, and 18-month-olds understand false belief. The findings have come from looking-time research methods (Onishi & Baillargeon, 2005; Scott & Baillargeon, 2009; Surian, Caldi, & Sperber, 2007), from examinations of anticipatory looking in eye-tracking paradigms (Southgate, Senju & Csibra, 2007), and from 18-month-olds' responses in active-interactive paradigms (Buttelman, Rickards, & Bortoli, 2009; Southgate, Chevallier & Csibra, 2010). In this chapter, we will not tackle this accumulating body of infant "false belief" research, but concentrate instead on preschool developments. We adopt this focus for (at least) three reasons. First the full, correct interpretation of the infant "false belief" data is not yet clear (e.g. Sodian, 2011). Some impressive cognition is going on, and the "classic" consensus that all the progress in understanding false belief emerges in the preschool years and not earlier clearly must be revised. Secondly, relatedly, it is not clear how the data from infancy and those in preschool fit together. The preschool data tap conceptions that are considerably more explicit and aware than the infant data, but contrasting explicit and implicit knowledge may not be the best way to talk about or reconcile the two sets of findings (e.g. Apperly & Butterfill, 2009). One thing that is clear is that the infant and preschool findings do fit together in some manner. Research now shows that infant competence in social cognition on looking-time tasks predicts later preschool competence in standard tasks (Kristen, Thoermer, Hofer, Aschersleben, & Sodian, 2006; Wellman, Phillips, Dunphy-Lelii, & LaLonde, 2004, Wellman, Lopez-Duran, LaBounty, & Hamilton, 2008; Yamaguchi, Kuhlmeier, Wynn, & VanMarle, 2009).

A final, probably most important reason to continue to concern ourselves with the preschool data is that the preschool conceptual developments are importantly related to children's everyday life, actions, and interactions. How so? Consider the data in Box 4.2 again. Amid strong common trends there is some obvious variation in preschool false-belief achievement across countries and, not explicitly evident in Box 4.2, also across individuals. Although almost all typically developing children master false belief before middle childhood, some children in some places come to this understanding earlier and some later. This variation helps researchers confirm the impact of achieving preschool theory-of-mind understandings. Children's performance on false-belief tasks is just one marker of these understandings, but differences in false belief understanding alone

Box 4.2



Explicit false belief tasks have children reason about an agent whose actions should be controlled by a false belief. A common task employs a change in locations, as depicted above. The child (not shown above) sees the character, Jill, put her chocolate in one location. The character leaves and while she cannot see, the chocolate gets moved. When the character returns the child is asked “Where will she look for her chocolate?” or “Where does she think her chocolate is?” Older children (5 years and over in many studies) answer correctly, like adults. Younger children answer incorrectly. They are not just random; they consistently say the agent will search in the new location (where it really is). Note that the task taps more than just attribution of ignorance (Jill doesn’t know where her chocolate is); rather it assesses attribution of false belief (Jill thinks—falsely—that her chocolate is in the drawer).



As shown in the graph above, children in different cultural-linguistic communities achieve false belief understanding more quickly or more slowly, yet in all locales they evidence the same trajectory—from below chance to above-chance performance (0 = chance in this graph) typically in the preschool years (Combined data from Wellman, Cross, & Watson, 2001 and Lin, Wellman, Tardif, & Sabbagh, 2008).

predict how, and how much, preschool children talk about people in everyday conversation (e.g. Dunn, 1996; Dunn & Brophy, 2005), their engagement in social pretend play (e.g. Astington & Jenkins, 1999), their social interactional skills (Lalonde & Chandler, 1995; Peterson, Slaughter & Paynter, 2007), including their attempts at persuasion and their participation in games like hide and seek (Bartsch, London & Campbell, 2007; Peskin & Ardino, 2003), and their interactions with and popularity with peers (e.g. Diesendruck & Ben-Eliyahu 2006; Siegal & Peterson, 2002; Slaughter, Dennis, & Pritchard, 2002; Watson, Nixon, Wilson, & Capage, 1999). These findings importantly confirm theory of mind's real-life relevance; moreover, they demonstrate that something definite and important is happening in children's theory of mind understandings in the preschool years.

In light of the emerging infant data, one theory that has been advanced about these preschool developments is that infant theory of mind defines the competence, and it is just the *expression* of that competence that is revealed in the preschool years. Early competence is masked in young preschool children in much research because of the executive function demands of standard preschool theory of mind tasks (e.g. Luo & Baillargeon 2010; Scholl & Leslie, 2001). On the strongest claim, that would make the preschool developments, more or less, nothing but executive-function development.

However, thinking of preschool false-belief tasks as just proxies for executive-function development cannot be the proper account, not the full nor most important story. First, false belief still significantly predicts aspects of children's conversations, their social interactional skills, their engagement in pretense, their interactions with and popularity with peers, their participation in games like hide and seek, in studies where executive functioning is controlled and factored out (e.g. Peskin & Ardino, 2003; Razza & Blair, 2008). Cross-cultural data are telling as well. Suppose for a moment that false belief performance does represent just executive-function achievement. Those children who early achieve adequate executive functioning should equally attain early good false belief performance. Intriguingly, there is evidence that in East Asia (e.g. Oh & Lewis, 2008), and specifically in China (Sabbagh, Xu, Carlson, Moses, & Lee, 2006), children have earlier developing executive-function skills relative to their Western peers (probably because parents and teachers place particular emphasis on the socialization of self-control). This earlier competence at executive-function, however, does *not* translate into better or earlier false belief understanding. In precise comparisons between preschoolers in Beijing (Sabbagh et al., 2006) and the USA (Carlson & Moses, 2001), Chinese children were consistently and significantly advanced in executive-functions (on eight different executive-function tasks), and yet at the same time for the same children there were no theory of mind differences between the Chinese and US children at 3½, 4, or 4½ years on four different "standard" preschool false belief tasks. (See Liu et al., 2008, for related US–Chinese comparisons.)

So our point for this chapter is that whatever is happening in infancy, genuine and important theory of mind progressions are occurring in the preschool years; conceptual progressions that are not just proxies for executive functions or general cognitive complexity and that impact children's social actions and lives. Moreover, returning to our insistence that understanding of theory of mind will be best advanced by more richly developmental data, preschool data now provide the best look at how theory of mind progresses developmentally.

Scaling theory of mind progressions

On this point, the preschool false-belief data alone (as in Box 4.2) provide an intriguing initial look at childhood development of theory of mind, as well as an intriguing initial demonstration of universality in childhood development of social cognition. However, although data like these

are standard for research in cognitive development, they are not very developmental. They show developmental progress with age, but it is really only passing or failing one task averaged across age. While the data show some universality (and some variability), it is not clear from false-belief data alone how to best understand these cross-national comparisons. Newer data unpack universality and variability more clearly and do so by clarifying more extended developmental progressions.

Consider these related things a child could know (or not know) about persons and minds: (a) people can have different desires for the same thing (diverse desires, or DD), (b) people can have different beliefs about the same situation (diverse beliefs, DB), (c) something can be true, but someone might not know that (knowledge access, KA), (d) something can be true, but someone might falsely believe something different (false belief, FB), (e) someone can feel one way but display a different emotion (hidden emotion, HE). These notions capture aspects of mental subjectivity, albeit different aspects (including mind–mind, mind–world, and mind–action distinctions). Listing them in this manner suggests that one could devise a set of tasks all with similar formats and procedures, pretty much like standard false belief tasks, for example, and see how children do. Several studies have now done just that.

Studies using such a battery of tasks, encompassing more than 500 preschoolers in the USA, Canada, Australia, and Germany, evidence a clear and consistent order of difficulty. It is the order listed above, with diverse desires easiest and hidden emotions hardest. For shorthand, let us call this sequence, DD>DB>KA>FB>HE. This sequence is highly replicable and significant—80% of these children show this pattern (Peterson, Wellman, & Liu, 2005; Peterson & Wellman, 2009; Wellman & Liu, 2004). Furthermore, longitudinal data for US preschoolers over the age range from 3 to 6 years confirm that individual children develop according to the same sequence, and at the same pace, as shown in the cross-sectional data (Wellman et al., 2011).

Culture and variation

So, these tasks—constituting a Theory of Mind Scale—reveal a robust sequence of understandings. What accounts for the consistency of sequence demonstrated so far? Clearly, a consistent sequence could result from innately programmed maturations. (Or similarly, it could result from maturationally unfolding gains in basic cognitive processes, say increases in executive function or in cognitive capacity.) Alternatively, a consistent sequence might result from processes of conceptual learning in which initial conceptions lead to later conceptions, shaped by relevant information and experiences. Crucially, if they are more shaped by relevant information and experiences then, in principle, sequences could be very different across children and groups.

Additional cross-cultural research, for example in China, addresses these possibilities. Assume that theory-of-mind understandings *are* the products of social and conversational experiences that vary from one community to another. Western and Chinese childhood experiences could be crucially different. Various authors have described an Asian focus on persons as sharing group commonalities and interdependence and a contrasting Western focus on persons as distinctively individual and independent (e.g. Markus & Kitiyama, 1991; Nisbett, 2003). These differences include differing emphases on common knowledge and perspectives vs. diversity of individual beliefs and perspectives. Moreover, Western and Chinese adults seem to manifest very different everyday epistemologies. Everyday Western epistemology is focused on truth, subjectivity, and belief; Confucian-Chinese epistemology focuses more on pragmatic knowledge acquisition and the consensual knowledge that all right-minded persons should learn (Nisbett, 2003; Li, 2001). Indeed, in conversations with young children, Chinese parents comment predominantly on “knowing”, whereas US parents comment more on “thinking” (Bartsch & Wellman, 1995; Tardif & Wellman, 2000).

In accord with such conversational-cultural preferences for emphasizing knowledge acquisition vs. belief differences, Chinese preschoolers evidence a consistent but different theory of mind sequence where KA and DB are reversed: DD>KA>DB>FB>HE (Wellman, Fang, Liu, Zhu, & Liu, 2006; Wellman et al., 2011). Both Western and Chinese children first understand basic aspects of desire (DD). However, the cultures diverge at the next step. Most Western children first appreciate belief differences (DB) and then acquisition of knowledge (KA). For Chinese children in Beijing, the ordering of these two steps is reversed. Most of them pass KA before DB.

This is not some singular peculiarity of Chinese mind and development; the same alternative sequence appears in Iranian preschool children (Shahaeian, Peterson, Slaughter, & Wellman, 2011). Despite profound differences in Iran's Muslim traditions and beliefs in contrast to Chinese Confucian/Buddhist/Communist ones, both China and Iran share collectivist family values emphasizing consensual learning, knowledge acquisition, respect for the wisdom of elders, and low tolerance for children's assertions of disagreement or independent belief. As a consequence, parents in both Iran and China often resemble each other and differ from Western parents in their socialization practices, and parental goals and values. Prototypically, Iranian and Chinese parents place stronger emphasis on children's conformity to tradition and emulating knowledgeable adults to overcome their ignorance. Western parents are correspondingly apt to more strongly encourage their children's thinking independently, listening to the views of peers, and freely and assertively expressing their own opinions. Intriguingly, data on children's progressive development through the Theory of Mind (ToM) scale are in line with these cultural variations in parental beliefs and practices. Most children in the USA and Australia form initial conceptualizations of mind in terms of differences of opinion, thus explaining their early mastery of the DB task. Children in Iran (Shahaeian et al., 2011), like their peers in China (Wellman et al., 2006), most often construct their initial awareness of thinking around the idea of acquiring knowledge, thus explaining their relatively earlier mastery of the KA task. Perhaps because their socialization confers less exposure to opinion diversity, they are likewise correspondingly slower than their Australian or US peers to master DB, even though overall rates of developmental progress through the full five steps of the ToM Scale are equivalent in all four countries.

We believe that sequence similarities from one culture to the next, coupled with cross-cultural differences like these, are especially important and revealing. For example, a developmental scale, encompassing a sequences of acquisitions achieved over a range of ages, can provide more informative comparisons of different children's understandings for use in individual differences comparisons. Similarly, a scale can help overcome a problem endemic to cross-cultural comparisons, namely how to validly compare children across different countries and communities. Often this is done by comparing two samples of convenience (e.g. one in the USA and one in China) on a single task (e.g. false belief performance). Yet two such samples, even if carefully matched for comparable ages, differ so widely (e.g. in languages they acquire, family experiences, nature and onset of school or preschool experiences) that evidence that one group is better or worse on a single task is almost impossible to interpret. When extended progressions are the same or different (e.g. revealing sequence differences) comparisons are considerably more informative and interpretable.

Of course, sequences are not the only issue—developmental timetables also matter. How long can and do these “preschool” developments take? How can differences in developmental timing be explained?

Atypical development: deafness

The false belief data in Box 4.2 already show that timetables can vary; some preschool children are quicker, some slower to achieve false belief understanding. However, in the bigger picture this may

not represent all that much variation. Pretty much everywhere, children achieve a similar understanding of false belief before the end of their preschool years. A similar suspicion might arise for the sequence data as well: sequence differences are intriguing, but actually, children proceed through more or less the same steps, and at more or less the same time—all within the bounds of the preschool period. Tightly restricted—not identical, but restricted—timetables might well reflect development of theory-of-mind understandings as largely under maturational control. If early progressive theory-of-mind understandings are built one upon the next; however, shaped by relevant information and experience, such a developmental process should be able to produce very different timetables.

It has been known for a long time that false belief understanding *is* seriously (not modestly) delayed in children with autism. This classic example has been well reported in prior editions of this volume. Most adolescents and adults with autism perform poorly on false belief tasks even after making allowances for possible deficits in verbal or non-verbal IQ (Happé, 1995; Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998). However, an autism diagnosis is replete with other known developmental atypicalities including neurological impairments, genetic abnormalities, and frequent general, across-the-board, cognitive impairment and delays. Autism could certainly have its own delayed maturational timetable. As we noted earlier, we argue for a special focus on deaf children.

Even when carefully selected so as to be free of all disabilities apart from hearing loss (including free from the social deficits, neurological impairments, and intellectual delays that are often associated with autism), deaf children of hearing parents (but *not* deaf children of deaf parents) are substantially delayed in understanding false belief (see Peterson, 2009; Peterson & Siegal, 2000; Siegal & Peterson, 2008, for reviews). Indeed they typically score no higher than children with autism of similar age and intellectual ability (e.g. Peterson, 2002; Peterson, Wellman & Liu, 2005). Studies of more than 700 severely or profoundly deaf offspring of hearing parents from several different countries (and hence different signed and spoken languages, and approaches to deaf education) consistently reveal predominant failure on false belief tasks throughout primary school (ages 6–12). This is in clear contrast to hearing preschoolers' predominant success by age 4 or 5. Even in high school, two cross-sectional studies of deaf teens from hearing families revealed pass rates of only 53 and 60%, respectively, at ages 13–17 (Edmondson, 2006; Russell et al., 1998). These serious delays apply just as much to deaf children who sign vs. communicate orally and, within the latter group, to those who use cochlear implants vs. external hearing aids (see Peterson, 2004, for a review).

An intriguing 5–10% percent minority of deaf children are native signers with signing deaf parents. Their theory of mind performance is in sharp contrast to that of matched deaf groups from hearing families. Thus, native signers in primary school consistently outperform their deaf classmates and equal their hearing ones (e.g. Courtin & Melot, 2000; Peterson & Siegal, 1999; Meristo, Falkman, Hjelmquist, Tedoldi, Surian, & Siegal, 2007; Schick, De Villiers, De Villiers, & Hoffmeister, 2007). When a parent is a deaf signer, the natively signing deaf child has ordinary conversational experiences from birth—albeit in sign language—unlike the case if there are hearing parents of deaf children who, despite extensive and conscientious efforts, almost never become fully proficient in sign language. Very few can communicate effectively with their deaf offspring in sign about unobservables like thoughts, feelings and other mental states (e.g. Moeller & Schick, 2006; Vaccari & Marschark, 1997). Furthermore, carefully controlled studies show that this difference in the early family conversational environment is directly linked with the native signer's superiority on theory-of-mind tests. Native signers outperform their deaf peers from hearing families even after statistically controlling for other potentially contributing factors like executive function

or current levels of lexical or syntactic language skill (e.g. Schick et al., 2007; Woolfe, Want & Siegal, 2002). In these ways, delayed theory of mind (specifically, for now, delayed false-belief understanding) is demonstrably not a direct consequence of deafness per se. Rather, it requires deafness in conjunction with upbringing in conversational situations that are reduced and problematic, as in purely hearing families and/or purely oral schools. Native signers' early opportunities to share, via sign, in conversations about thoughts and feelings at home and at school may well be crucial to their timely acquisition of false belief understanding. "No other social activity requires more negotiation with other minds than conversation" (Malle & Hodges, 2005; p. 5).

Again, however, a focus on false belief alone is limiting. More informatively, when deaf children of hearing parents receive the Theory of Mind Scale they too evidence a consistent sequence of progression, but one that is delayed at every step of the way (e.g. Peterson & Wellman, 2009; Peterson et al., 2005). It often takes deaf children 10–12 years, or more, to progressively achieve what hearing children (and deaf children of deaf parents) progressively achieve in 4–6 years (Peterson, 2009; Wellman et al., 2011). In ToM Scale data summed across various studies and across 66 deaf children of hearing parents, the average age of acquisition for DB was 8 years, for KA it was 10 years, for FB it was 11 ½ years, and for HE (Hidden Emotion), 12 ½ years (Wellman et al., 2011).

Furthermore, longitudinal data (Wellman et al., 2011) confirm that the same ToM Scale sequence (DD>DB>KA>FB>HE) that characterizes the development of deaf children cross-sectionally also accurately describes the development longitudinally for individual deaf children as they progress through primary school and into high school. Figure 4.1 captures that data. Like the cross-sectional data, these longitudinal data confirm how seriously delayed deaf children of hearing parents are in developing theory of mind understandings. On average, deaf children (beginning at age 8) take 41 months to progress longitudinally through the scale as far as a hearing 3-year-old progresses in just 12 months.

The performance of these deaf children (as in Figure 4.1) also speaks strongly against any maturational, critical-period analysis of theory of mind. According to a critical period account, deprivation of some crucial input or experience at some specified early age (e.g. the preschool years: Siegal & Varley 2002; or the period before age 12: Morgan & Kegal, 2006) will result in a permanent difficulty or deficit in theory-of-mind development, no matter how richly stimulating the environment may become subsequent to that critical time. It is difficult to rule out such a hypothesis with cross-sectional evidence. Yet these newer longitudinal data, by revealing steady progress by deaf children both in false belief understanding (e.g. Peterson, 2009) and on the theory of mind scale (as shown in Figure 4.1) through primary school and into high school are inconsistent with this critical period view.

Furthermore, a provocative longitudinal study of a unique group of Nicaraguan deaf adults (Pyers & Senghas, 2009) longitudinally followed two cohorts. The first, throughout their growing up, had been limited to using a pidgin form of signing that had no terms for thinking or other cognitive states. Not surprisingly, they continued to fail FB tests even in adulthood, but then, while in their late twenties, they began to interact at a local deaf club with adults from the second cohort. The latter had, through interacting together in primary and high school, created a sign language equipped with cognitive terms and the first-cohort adults eventually learned it from them. A longitudinal test after 2 years of this informal conversational contact revealed dramatic theory of mind gains for the older, language-deprived cohort who now equaled the second cohort (most of whom had mastered false belief by their early to mid teens). Together with longitudinal evidence from late-signing deaf children, these results demonstrate that first-time mastery of "preschool" theory of mind understanding is possible well beyond the bounds of any postulated critical periods and, indeed, well into adulthood.

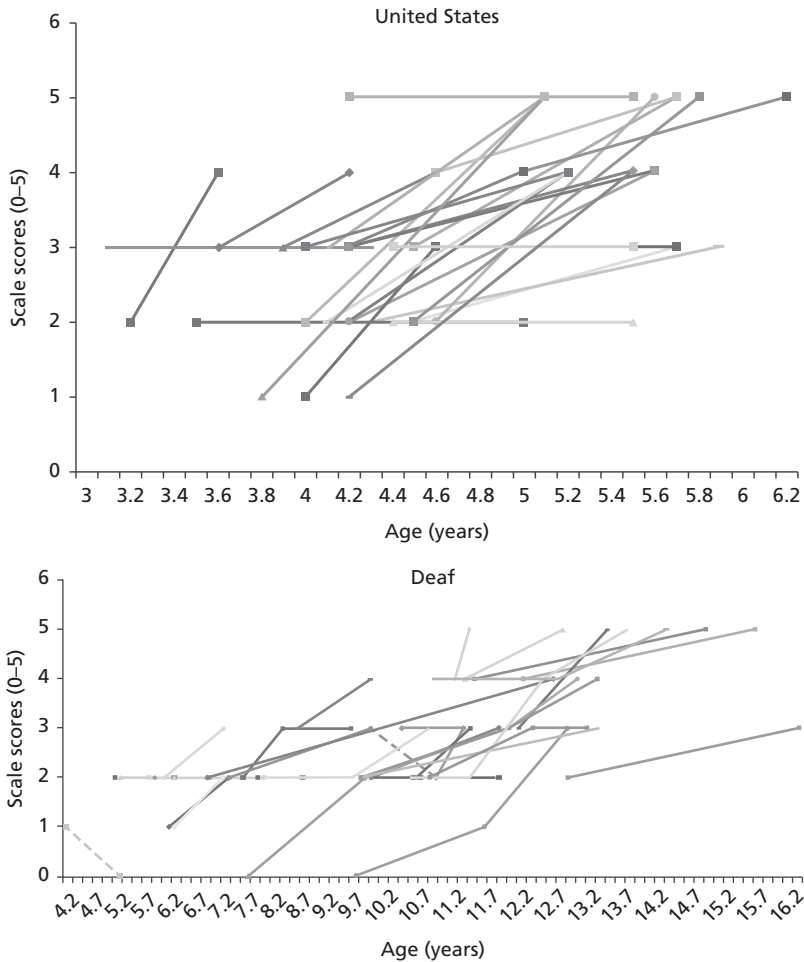


Figure 4.1 Panels showing longitudinal changes on the ToM Scale. Each gray-tone line shows the longitudinal trajectory for a single child. Figures reproduced from Child Development 82 (3), Henry M. Wellman, Fuxi Fang, and Candida C. Peterson, Sequential Progressions in a Theory-of-Mind Scale: Longitudinal Perspectives, pp, 780–92 © 2011 The Society for Research in Child Development, Inc. See also Plate 1.

At the same time, all the evidence clearly shows that most deaf children of hearing parents are seriously delayed in theory of mind understandings, demonstrating the extent to which socio-cultural-conversational interchanges affect children's theory of mind growth. These data as to dramatic timetable differences mean that it is worth readdressing the issue of theory of mind sequences for deaf children of hearing parents. If the differences apparent in Chinese or Iranian vs. US and Australian input and experiences can lead to differences in sequences in the hearing preschool child's case, then the far more striking differences in conversational input and socialization experiences that exist between hearing and deaf children should affect theory of mind sequences, as well as timetables. Deaf children provide an example of children growing up in quite different socio-linguistic, socio-interactive environments in comparison to their hearing peers, cultural differences that are arguably at least as wide as those between US and Chinese children. Accounts based on construing theory of mind development as based on conceptual learning from socially

variable experiences and inputs should predict that sequences will vary between these groups (at least in some relevant ways).

In a further study we probed the prior finding of identical sequences more deeply by adding a focus on children's understanding of social pretense (Petersen & Wellman, 2009). Why this focus? Briefly, comparisons and sequences between understanding pretense and belief are theoretically intriguing (much like the comparisons between knowledge and belief). Moreover, deaf and hearing children's experiences of pretend play are likely to be quite different, providing an important test case for examining similar or different sequences of understanding.

Understanding pretense reflects an understanding of others' mental states, at least for adults and older children (Harris, 2005; Harris & Kavanaugh, 1993; Lillard, 1993; Richert & Lillard, 2002). Yet conceptually, pretense and belief clearly differ. Beliefs are "supposed" to be accurate. In other words, they are accepted not just as representations, but as representations of factual reality. While beliefs can be false, in general they are meant to be true. In contrast, in pretense truth is less an issue. Indeed, the whole point of pretense is to create an imaginary representational situation that departs from the truth of present reality. In line with this analysis, understanding pretense (and imagination) has been shown to be easier than understanding false belief in several studies with hearing children (Custer, 1996; Gopnik & Slaughter, 1991; Hickling, Wellman, & Gottfried, 1997—but see Lillard 1993 and Richert & Lillard 2002 for arguments and data that at least some forms of pretense understanding are harder and later-developing than understanding of false beliefs).

Pretense additionally stands out as important because childhood engagement in pretend play relates to and may influence typically developing children's theory of mind understandings. Thus, preschoolers' false belief scores are often found to correlate with their frequency of engaging in pretend play (Taylor & Carlson, 1997; Youngblade & Dunn, 1995), at least when it is socially shared (e.g. Astington & Jenkins, 1999), and the well-established sibling theory of mind advantage (whereby preschoolers with child-aged siblings master false belief at younger ages than only-children) is often ascribed to added opportunities for make believe play in a sibling family (e.g. Perner, Ruffman & Leekam, 1994). Much less is known about these relations and influences for deaf children, but if social experiences crucially influence pretense understanding and theory of mind development, there are reasons to expect that the experience of pretense, and thus the representational understanding of it as a mental state, might be quite different for deaf children. One can easily imagine two distinct scenarios.

On the one hand, it is clear that deaf children (in hearing families) are generally delayed in their pretense actions and interactions (e.g. Brown et al., 2001), just as they are generally delayed in their understanding of mental states such as beliefs and false beliefs. Such data suggest that (especially if early shared pretend interactions are key) deaf children may be particularly delayed in understanding (as well as participating in) pretense. Preschoolers who are severely or profoundly deaf may miss out on pretense discourse and other experiences with social pretending owing to lack of a common language (speech or sign) that they can fluently share with hearing family members or their deaf peers.

On the other hand, granting overall delays in theory of mind understanding, pretense understanding and experiences may be less delayed or less impaired in deaf children than understandings of belief. For example, for deaf children, sharing pretense stipulations with others may proceed in largely nonverbal ways via gesture, pantomime, or toy manipulation. Indeed, simple pretense stipulations might arguably occur just as easily via gestures as via words (e.g. by holding a banana to one's ear or pressing imaginary keys to simulate a mobile phone). This could make non-verbal gestures (a strength of deaf children) a facilitative medium for the social sharing of mental states with parents and playmates, within pretense.

We addressed these possibilities in research including a focus on pretense understanding along with other aspects of a developing theory of mind (Peterson & Wellman, 2009). Specifically, we added a test of pretense understanding (PU) to the battery of theory of mind tasks in the ToM Scale described above, namely; DD, DB, KA, FB, HE. For hearing preschoolers, we found that our pretense understanding task (PU) scaled consistently as an intermediate point between the ToM Scale concepts of knowledge access and false belief (i.e., DD>DB>KA>**PU**>FB>HE). For deaf children of hearing parents, however, the pattern was different. Although their pretense understanding, just like hearing children's, scaled predictably as a statistically reliable component of the full ToM Scale, and although it emerged at a later age than hearing children's, it consistently occupied an earlier scale step. For deaf children pretense preceded both KA and FB (i.e. their sequence was: DD>DB>**PU**> KA>FB>HE). This, variation in sequence (similar to the Western vs. Chinese/Iranian sequence contrast for DB and KA as described above) suggests a further role for varied social and conversational experience in the sequences through which understandings of mind emerge.

Continuing the sequence into late childhood and beyond

As the findings reviewed thus far from deaf children and adolescents clearly illustrate, continuing ToM development is possible well beyond childhood. Indeed, the same is true for typically developing children, adolescents and adults. Dramatic as the preschool gains are (outlined in our earlier overview of typical preschoolers' theory of mind development), still older children develop additional understandings, including increasingly reflective ideas about minds, brains, and mental life (e.g. Wellman & Johnston, 2008). To illustrate, children's understanding of thinking shows considerable development. While 3- and 4-year-olds know that thinking is an internal mental event (different from looking, talking, or touching) and that the contents of one's thoughts (e.g. a thought about a dog) are not public or tangible (e.g. Wellman & Estes, 1986; Wellman, Hollander & Schult, 1996; Richert & Harris, 2006), they fail to recognize the constant flow of ideas and thoughts experienced in everyday life and involved in actively, consciously thinking. Thus, 7-year-olds and adults assert that a person sitting quietly with blank expression is still experiencing "some thoughts and ideas" and that it is nearly impossible to have a mind completely "empty of thoughts and ideas"; but children 5 and younger do not share these intuitions (Flavell, Green, & Flavell, 1993, 1995). Coming to recognize that thinking streams along constantly is a development that follows on the heels of initial preschool understandings of mental contents.

Another set of achievements that nicely illustrates older children's increasingly reflective social cognition concerns their understanding of nonliteral communication, such as metaphor (Wellman & Hickling, 1994), sarcasm (Happé, 1993; Filippova & Astington, 2010), and many forms of joking (Winner, 1993; Winner & Gardener, 1993; Winner & Leekam, 1991) and teasing (Dunn, 1996; Dunn & Brophy, 2005; Dunn & Hughes, 1998). Such conversational situations use language to express meanings that are literally false, and require that the listener appreciate the differences between beliefs, communicative intentions, and messages. When a speaker comments "It's great weather today!" in the midst of cold and pouring rain, most children aged 5–7 fail to perceive that any sarcasm is intended (e.g. Filippova & Astington, 2008). Some understanding of simple false-naming jokes ("look a shoe", said of a hat) is evident for typical 2-year-olds (Baron-Cohen, 1997), however, with increasing age and exposure to varied conversational and social experiences with parents, siblings and peers (e.g. Recchia, Howe, Ross & Alexander, 2010), typically-developing children gradually come to further understand the interplay between mind, meaning, and occurrences, including the awareness of how speakers use irony and sarcasm to convey meanings opposite to

the literal meaning of their words. Numerous studies, therefore, show a steady improvement in the awareness of nonliteral meanings of ironic and sarcastic utterances like this, through age 10 to 12 (e.g. Demorest et al., 1991; Filippova & Astington, 2008, 2010; Pexman & Glenwright, 2007). Indeed, “progress in understanding the social-cognitive functions of irony ... continues well into adulthood” (Filippova & Astington, 2010, p. 218).

Early difficulty with sarcasm and irony is evident not only in typical development, but also in adult and child populations with autism (e.g. Happé, 1993), schizophrenia (Langdon & Coltheart, 2004) or acquired brain damage (e.g. Gallagher, Happé, Brunswick, Fletcher, Frith & Frith, 2000). More relevant to our focus is that deaf educators and parents of deaf individuals report continuing difficulties with non-literal language as deaf children develop into adolescents and adults. Gregory, Bishop and Sheldon’s (1995) interviews with the hearing parents of deaf young adults revealed persistent difficulties with non-literal language and sarcastic humour even among those who were functioning quite successfully both as mature communicators (in speech or sign) and in everyday life within their communities. For example, one hearing mother reported that her 19-year-old daughter, a British Sign Language (BSL) user, “doesn’t know the meaning of a joke; if you say something, it’s serious. She can’t see a double meaning ... as far as language goes, you can’t play around with it” (p. 33). In general, verbal humour and sarcasm posed problems for nearly 60% of this sample of severely or profoundly deaf young adults, with no distinction between signers (of BSL or signed English) and oral-language users.

Sarcasm understanding is thus a sophisticated developmental achievement (e.g. Baron-Cohen, 2000; Rajendran, Mitchell, & Rickards, 2005; White, Hill, Happé, & Frith, 2009) and one with well-established empirical links to theory of mind concepts like false belief (e.g. Filippova & Astington, 2008). Therefore, in recent research (Peterson, Wellman & Slaughter, 2012) we began with the well-validated ToM Scale (DD>DB>KA>FB>HE) and created a sarcasm detection task identical in style and question format to other scale tasks. Even after controlling statistically for variations in linguistic ability, results showed that sarcasm understanding was more difficult than the final step on the original scale (HE) for typically developing children and for those with deafness (and autism). Guttman scaling analyses confirmed that the understanding of the discrepancy between spoken communicative intent and literal word meaning, as assessed by our sarcasm task, is a reliably more advanced theory of mind achievement and one that, even in typical development begins, but is not yet completed during middle childhood. Note that sarcasm detection scaled as a sixth step in the developmental progression of theory of mind, and did so similarly for hearing and deaf children, but at substantially different times. Sarcasm detection was especially difficult for deaf children even in comparison to age-matched peers with autism. We take such data to provide further support for social conversational accounts of theory of mind development by highlighting the value of children’s varied participation in everyday social exchanges—ranging from pretend play, teasing, and joking, to emotional concealment, sarcasm, and other affectively-laden non-literal uses of language—for fostering the timely achievement of mature social understanding. Such experiences and achievements are part and parcel of mature social interaction for typically developing children and they apparently pose extended social difficulties for theory-of-mind-delayed deaf children of hearing parents (as well as for language-delayed children with autism).

Conclusions

The development of a theory of mind, or an everyday understanding of people’s behavior in terms of what they know, think, intend and hope, begins at birth and follows a trajectory of striking and progressive developmental gains throughout childhood and into adolescence. Recent research

clearly shows that infants as young as 12–18 months are already capable of some remarkable insights into the subjectivity of human mental experience. They ascribe intentionality to infer goals and desires even from failed or incomplete actions. In perceptual situations, they can often distinguish their own awareness from someone else's lack of it.

Yet it is equally apparent that the developmental story does not end in infancy. The vast majority of 3-year-olds routinely fail simple challenges to their ToM understanding both in the laboratory (as on false belief tests) and in everyday life (as when playing hide-and-seek, keeping secrets, discerning lies, negotiating pretense, or interacting skillfully with peers). Yet, by age 5 or 6, success on these everyday and laboratory applications of theory of mind is consistent and widespread across cultures and methodologies (see Box 4.2).

Nor does theory of mind development stop at the end of preschool. As older children and adolescents gain increasing social and conversational experiences in the classroom and on the playground, their understanding of mental life becomes ever more nuanced, reflective, versatile and complex. They gain new insights into cognition and the stream of consciousness, including a deeper appreciation of the subjectivity, interpretivity and diversity of thought, memory and belief. Their understanding of emotion grows as does their appreciation and use of language and its pragmatics, including nonliteral communications like sarcasm. These more advanced insights are not simply a haphazard or piecemeal collection of new ideas. Just like the scalable progressions arising from late infancy to the end of preschool, these further, more advanced manifestations of understanding other minds include reliably sequential developments that follow, elaborate, and build upon the traditional theory-of-mind hallmark, namely success on explicit, inferential false belief tests. Thus, recent cross-sectional and longitudinal scaling studies across several different cultures have extended the understanding of theory of mind development by documenting a reliable progression of conceptual achievements over five or six sequential steps, beginning with diverse desires (toddlers' understanding that different people want different things) and continuing through preschoolers' awareness of diverse beliefs, knowledge access and false belief to older children's awareness of the subtle socio-cognitive underpinnings of hidden emotion and sarcastic communication.

Despite broad consistency in these extended developmental sequences, there are also variations in developmental patterns for individual children and cultural groups that testify clearly to the influences of particular inputs and interactive social experiences on the progression of theory of mind development. Comparison between children in Western cultures (e.g. Australia, Germany, and the USA), in contrast to those in China and Iran nicely illustrate these differences, differences that are predictable based on contrasts between Western and Eastern/Middle-Eastern parenting beliefs, values and socialization practices.

Of course, not all children develop theory of mind on the same early timetable. Our focal example of theory-of-mind delays has been deaf children of hearing parents. These examples from deafness highlight several important conclusions about theory of mind development for children generally. For example, one idea that cannot readily be tested with typically developing children (owing to their universally rapid theory of mind mastery) is whether there is a critical period for theory of mind achievements. Recent longitudinal research with deaf children and adolescents (see Figure 4.1) and adults (Pyers & Senghas, 2009), effectively dispels the critical period idea for "preschool" theory of mind insights.

Data on deaf children's developmental theory-of-mind sequences and timetables have additional theoretical implications. To illustrate, extended theory-of-mind scaling research has revealed that dealing with nonliteral language, as reflected in appreciation of sarcasm, is a consistently late or advanced theory of mind understanding for all groups of children tested so far. Yet it is unusually

delayed for deaf children. This makes sense in terms of conversational input and social interaction experience: Sarcasm is poorly understood, and hence rarely used, even by deaf adults (e.g. Gregory, Sheldon, & Bishop, 1995).

To conclude, exciting new horizons for theory of mind research go beyond developments in infancy to encompass preschoolers, older children, and adults. Deeper attention to development, coupled with new developmental methods, have established new insights as well as revealing promising directions for future research. Research with deaf children, in particular, combines with that from hearing populations and from children with autism to provide new insights that promise additional theoretical and applied advances for comprehending children's understanding of other minds.

References

- Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4), 953–70.
- Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental Psychology*, 35(5), 1311–20.
- Baldwin, D. A., & Moses, L. J. (1996). The ontogeny of social information gathering. *Child Development*, 67(5), 1915–39.
- Baron-Cohen, S. (1997). Hey! It was just a joke! Understanding propositions and propositional attitudes by normally developing children and children with autism. *Israel Journal of Psychiatry and Related Science*, 34(3), 174–8.
- Baron-Cohen, S. (2000). Theory of mind and autism. In S. Baron-Cohen, H. Tager-Flusberg & D. Cohen (Eds.), *Understanding Other Minds* (pp. 3–17). Cambridge: Cambridge University Press.
- Bartsch, K., London, K., & Campbell, M. D. (2007). Children's attention to beliefs in interactive persuasion tasks. *Developmental Psychology*, 43(1), 111–20.
- Bartsch, K., & Wellman, H. M. (1995). *Children Talk About the Mind*. New York: Oxford University Press.
- Brandone, A. C., & Wellman, H. M. (2009). You can't always get what you want: infants understand failed goal-directed actions. *Psychological Science*, 20(1), 85–91.
- Brown, P., Rickards, F. & Bortoli, A. (2001). Structures underpinning pretend play and word production in young children and children with hearing loss. *Journal of Deaf Studies and Deaf Education*, 6, 15–31.
- Butler, S., Caron, A., & Brooks, R. (2000). Infant understanding of the referential nature of looking. *Journal of Cognition and Development*, 1, 359–77.
- Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112(2), 337–42.
- Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, 72(4), 1032–53.
- Carpenter, M., Aktar, N., & Tomasello, M. (1998a). 14- to 18-month old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development*, 21, 315–30.
- Carpenter, M., Nagell, K., & Tomasello, M. (1998b). 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), i–vi, 1–143.
- Csibra, G., Gergely, G., Biro, S., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of “pure reason” in infancy. *Cognition*, 72, 237–67.
- Courtin, C., & Melot, A. M. (1998). Development of theories of mind in deaf children. In M. Marschark & D. M. Clark (Eds.), *Psychological Perspectives on Deafness* (pp. 79–102). Mahwah: Lawrence Erlbaum Associates, Inc.
- Custer, W. L. (1996). A comparison of young children's understanding of contradictory representations in pretense, memory, and belief. *Child Development*, 67(2), 678–88.

- 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: Cambridge University Press.
- Demorest, A., Silberstein, L., Gardner, H. & Winner, E. (1983). Telling it like it isn't. *British Journal of Developmental Psychology*, 1, 121–34.
- Diesendruck, G., & Ben-Eliyahu, A. (2006). The relationships among social cognition, peer acceptance, and social behavior in Israeli kindergarteners. *International Journal of Behavioral Development*, 30(2), 137–47.
- Dunphy-Lelii, S., & Wellman, H. M. (2004). Infants' understanding of occlusion of others' line-of-sight: Implications for an emerging theory of mind. *European Journal of Developmental Psychology*, 1, 49–66.
- Dunn, J. (1995). Children as psychologists: The later correlates of individual differences in understanding of emotions and other minds. *Cognition & Emotion*, 9, 187–201.
- Dunn, J. (1996). The Emanuel Miller Memorial Lecture 1995 Children's Relationships: Bridging the divide between cognitive and social development. *Journal of Child Psychology and Psychiatry*, 37, 507–18.
- Dunn, J. & Brophy, M. (2005). Communication, relationships, and individual differences in children's understanding of mind. In Astington, J. W. and Baird, J. A. (Eds), *Why Language Matters for Theory of Mind* (pp. 50–69). Oxford: Oxford University Press.
- Dunn, J. & Hughes, C. (1998). Young children's understanding of emotions within close relationships. *Cognition & Emotion*, 12, 171–90.
- Edmondson, P. (2006). Deaf children's understanding of other people's thought processes. *Educational Psychology in Practice*, 22(2), 159–69.
- Filippova, E., & Astington, J. W. (2008). Further development in social reasoning revealed in discourse irony understanding. *Child Development*, 79(1), 126–38.
- Filippova, E., & Astington, J. W. (2010). Children's understanding of social-cognitive and social-communicative aspects of discourse irony. *Child Development*, 81(3), 913–28.
- Flavell, J. H., Green, F. L., & Flavell, E. R. (1993). Children's understanding of the stream of consciousness. *Child Development*, 64, 387–98.
- Flavell, J. H., Green, F. L., & Flavell, E. R. (1995). Young children's knowledge of thinking. *Monographs of the Society for Research in Child Development*, 243 (entire serial), 60(1), 1–114.
- Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: Bradford Books/MIT Press.
- Gallagher, H., Happé, F., Brunswick, N., Fletcher, P., Frith U., & Frith, C. (2000). Reading the mind in cartoons and stories: An fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, 38, 11–21.
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Science*, 7, 287–92.
- Gopnik, A., & Slaughter, V. (1991). Young children's understanding of changes in their mental states. *Child Development*, 62(1), 98–110.
- Gregory, S., Sheldon, L., & Bishop, J. (1995). *Deaf young people and their families: Developing understanding*. Cambridge: Cambridge University Press.
- Hamlin, J. K., Hallinan, E. V., & Woodward, A. L. (2008). Do as I do: 7-month-old infants selectively reproduce others' goals. *Developmental Science*, 11, 487–94.
- Happé, F. (1993). Communicative competence and theory of mind in autism. *Cognition*, 48, 101–19.
- Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development*, 66(3), 843–55.
- 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.
- Harris, P. L. (2006). Social cognition. In W. Damon & R. M. Lerner (Eds), *Handbook of Child Psychology. Volume 2. Cognition, Perception, and Language*, Vol. 2 (pp. 811–58). Hoboken: John Wiley & Sons.

- Harris, P. L., & Kavanaugh, R. D. (1993). Young Children's understanding of pretense. *Monographs of the Society for Research in Child Development*, 58 (1), i–107.
- Hickling, A. K., Wellman, H. M., & Gottfried, G. (1997). Preschoolers' understanding of others' mental attitudes toward pretend happenings. *British Journal of Developmental Psychology*, 15, 339–54.
- Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child Development*, 76(2), 356–70.
- Kristen, S., Thoermer, C., Hofer, T., Aschersleben, G., & Sodian, B. (2006). Skalierung von "theory of mind" aufgaben (Scaling of theory of mind tasks). *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 38, 186–95.
- Lalonde, C. E., & Chandler, M. J. (1995). False belief understanding goes to school: On the social-emotional consequences of coming early or late to a first theory of mind. *Cognition and Emotion*, 9(2–3), 167–85.
- Langdon, R., & Coltheart, M. (2004). Recognition of metaphor and irony in young adults. *Psychiatry Research*, 125, 9–30.
- Li, J. (2001). Chinese conceptualization of learning. *Ethos*, 29(2), 111–37.
- Lillard, A. S. (1993). Young children's conceptualization of pretense: Action or mental representational state? *Child Development*, 64(2), 372–86.
- Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology*, 44(2), 523–31.
- Luo, Y., & Baillargeon, R. (2010). Toward a mentalistic account of early psychological reasoning. *Current Directions in Psychological Science*, 19(5), 301–7.
- Malle, B. & Hodges, S. (2005). *Other Minds: How Humans Bridge the Divide Between Self and Others*. New York: Guilford Press.
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224–53.
- Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31, 838–50.
- Meltzoff, A. N., & Brooks, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology*, 44(5), 1257–65.
- Meristo, M., Falkman, K., Hjelmquist, E., Tedoldi, M., Surian, L., & Siegal, M. (2007). Language access and theory of mind reasoning: Evidence from deaf children in bilingual and oral environments. *Developmental Psychology*, 43, 1156–69.
- Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, 78(2), 622–46.
- Moeller, M. P., & Schick, B. (2006). Relations between maternal input and theory of mind understanding in deaf children. *Child Development*, 77, 751–66.
- Moll, H., & Tomasello, M. (2004). 12- and 18-month-old infants follow gaze to spaces behind barriers. *Developmental Science*, 7, F1–9.
- Morgan, G., & Kegal, J. (2006). Nicaraguan sign language and theory of mind: The issue of critical periods and abilities. *Journal of Child Psychology & Psychiatry*, 47, 811–19.
- Moore, C., & Corkum, V. (1994). Social understanding at the end of the first year of life. *Developmental Review*, 14, 349–72.
- Nisbett, R. E. (2003). *The Geography of Thought: How Asians and Westerners Think Differently—and Why*. New York: Free Press.
- Oh, S., & Lewis, C. (2008). Korean preschoolers' advanced inhibitory control and its relation to other executive skills and mental state understanding. *Child Development*, 79(1), 80–99.

- Olineck, K. M., & Poulin-Dubois, D. (2005). Infants' ability to distinguish between intentional and accidental actions and its relation to internal state language. *Infancy*, 8, 91–100.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308, 255–8.
- Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: You catch it from your sibs. *Child Development*, 65(4), 1228–38.
- Peskin, J., & Ardino, V. (2003). Representing the mental world in children's social behavior: Playing hide-and-seek and keeping a secret. *Social Development*, 12(4), 496–512.
- Peterson, C. C. (2002). Drawing insights from pictures: The development of concepts of false drawing and false belief in children with deafness, normal hearing and autism. *Child Development*, 73, 1442–59.
- Peterson, C. C. (2004). Theory-of-mind development in oral deaf children with cochlear implants or conventional hearing aids. *Journal of Child Psychology and Psychiatry*, 45, 1–11.
- Peterson, C. C. (2009). Development of social-cognitive and communication skills in children born deaf. *Scandinavian Journal of Psychology*, 50(5), 475–83.
- Peterson, C. C., & Siegal, M. (1999). Representing inner worlds: Theory of mind in autistic, deaf and normal hearing children. *Psychological Science*, 10, 126–9.
- Peterson, C. C., & Siegal, M. (2000). Insights into a theory of mind from deafness and autism. *Mind & Language*, 15, 123–45.
- Peterson, C. C., Slaughter, V. & Paynter, J. (2007). Social maturity and theory-of-mind development in typically-developing children and those on the autism spectrum. *Journal of Child Psychology and Psychiatry*, 48, 1243–50.
- Peterson, C. C. & Wellman, H. M. (2009). From fancy to reason: Scaling deaf and hearing children's understanding of mind. *British Journal of Developmental Psychology*, 27, 297–310.
- Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development*, 76(2), 502–17.
- Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory of mind scales for typically developing children, and those with deafness, autism, or Asperger Syndrome. *Child Development*, 83(2), 469–85.
- Pexman, P. & Glenwright, M. (2007). How do typically developing children grasp the meaning of verbal irony? *Journal of Neurolinguistics*, 20, 178–96.
- Phillips, A. T., & Wellman, H. M. (2005). Infants' understanding of object-directed action. *Cognition*, 98(2), 137–55.
- Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science*, 20(7), 805–12.
- Rajendran, G., Mitchell, P., & Rickards, H. (2005). How do individuals with Asperger Syndrome respond to nonliteral language and inappropriate requests in computer-mediated communication? *Journal of Autism and Developmental Disorders*, 35, 429–43.
- Razza, R. & Blair, C. (2008). Associations among false-belief understanding, executive function and social competence. *Journal of Applied Developmental Psychology*, 30, 332–43.
- Recchia, H., Howe, N., Ross, H. & Alexander, S. (2010). Children's understanding and production of verbal irony in family conversations. *British Journal of Developmental Psychology*, 28, 255–74.
- Richert, R. A., & Lillard, A. S. (2002). Children's understanding of the knowledge prerequisites of drawing and pretending. *Developmental Psychology*, 38(6), 1004–15.
- Richert, R. A., & Harris, P. L. (2006). The ghost in my body: Children's developing concept of the soul. *Journal of Cognition and Culture*, 6(3–4), 409–27.
- Russell, P. A., Hosie, J. A., Gray, C. D., Scott, C., Hunter, N., Banks, J. S., et al. (1998). The development of theory of mind in deaf children. *Journal of Child Psychology and Psychiatry*, 39(6), 903–10.

- Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind: A comparison of Chinese and U. S. preschoolers. *Psychological Science*, 17, 74–81.
- Schick, B., De Villiers, P., De Villiers, J., & Hoffmeister, R. (2007). Language and theory of mind: A study of deaf children. *Child Development*, 78, 376–96.
- Scholl, B. & Leslie, A. (2001). Minds, modules and meta-analysis. *Child Development*, 72, 696–701.
- Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18months. *Child Development*, 80(4), 1172–96.
- Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2001). Culture and the sequence of steps in theory of mind development. *Developmental Psychology*, 47(5), 1239–47.
- Siegal, M. & Peterson, C. C. (2008). Language and theory of mind in atypically developing children: Evidence from studies of deafness, blindness, and autism. In C. Sharp, P. Fonagy & I. M. Goodyer (Eds), *Social Cognition and Developmental Psychopathology* (pp. 81–112). Oxford; New York: Oxford University Press.
- Siegal, M., & Varley, R. (2002). Neural systems involved in ‘theory of mind’. *Nature Reviews Neuroscience*, 3(6), 463–71.
- Slaughter, V., Dennis, M. J., & Pritchard, M. (2002). Theory of mind and peer acceptance in preschool children. *British Journal of Developmental Psychology*, 20(4), 545–64.
- Sodian, B. (2011). Theory of mind in infancy. *Child Development Perspectives*, 5, 39–43.
- Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others’ referential communication. *Developmental Science*, 13(6), 907–12.
- Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18(7), 587–92.
- Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18(7), 580–6.
- Tardif, T., & Wellman, H. M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology*, 36(1), 25–43.
- Taylor, M., & Carlson, S. M. (1997). The relation between individual differences in fantasy and theory of mind. *Child Development*, 68(3), 436–55.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5), 675–91.
- Tomasello, M., & Haberl, K. (2003). Understanding attention: 12- and 18-month-olds know what is new for other persons. *Developmental Psychology*, 39, 906–12.
- Vaccari, C., & Marschark, M. (1997). Communication between parents and deaf children: Implications for social-emotional development. *Journal of Child Psychology and Psychiatry*, 38, 793–801.
- Watson, A. C., Nixon, C. L., Wilson, A., & Capage, L. (1999). Social interaction skills and theory of mind in young children. *Developmental Psychology*, 35, 386–91.
- Wellman, H. M. (1990). *The Child’s Theory of Mind*. Cambridge: MIT Press.
- Wellman, H. M. (2011). Developing a theory of mind. In U. Goswami (Ed.), *The Blackwell Handbook of Childhood Cognitive Development*, 2nd edn (pp. 258–284). New York: Blackwell.
- Wellman, H. M., Cross, D., & Watson, J. (2001). A meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3), 655–84.
- Wellman, H. M., & Estes, D. (1986). Early understanding of mental entities: a reexamination of childhood realism. *Child Development*, 57(4), 910–23.
- Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science*, 17(12), 1075–81.
- Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory of mind scale: Longitudinal perspectives. *Child Development*, 82, 780–92.

- Wellman, H. M., & Hickling, A. K. (1994). The mind's "I": children's conception of the mind as an active agent. *Child Development*, 65(6), 1564–80.
- Wellman, H. M., Hollander, M., & Schult, C. A. (1996). Young children's understanding of thought bubbles and of thoughts. *Child Development*, 67(3), 768–88.
- Wellman, H. M., & Johnson, C. N. (2008). Developing dualism: From intuitive understanding to transcendental ideas. In A. Antonietti, A. Corradini & E. Lowe (Eds), *Psychophysical dualism today: An interdisciplinary approach* (pp. 3–35). Lanham: Lexington Books.
- Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75(2), 523–41.
- Wellman, H. M., Lopez-Duran, S., LaBounty, J., & Hamilton, B. (2008). Infant attention to intentional action predicts preschool theory of mind. *Developmental Psychology*, 44(2), 618–623.
- Wellman, H. M., Phillips, A. T., Dunphy-Lelii, S., & LaLonde, N. (2004). Infant social attention predicts preschool social cognition. *Developmental Science*, 7(3), 283–8.
- White, S., Hill, E., Happé, F. & Frith, U. (2009). The Strange Stories: Revealing mentalizing impairments in autism. *Child Development*, 80, 1097–117.
- Winner, E., & Gardner, H. (1993). Metaphor and irony: Two levels of understanding. In A. Ortony (Ed.), *Metaphor and Thought*, 2nd edn (pp. 425–43). New York: Cambridge University Press.
- Winner, E., & Leekam, S. (1991). Distinguishing irony from deception: Understanding the speaker's second-order intention. *British Journal of Developmental Psychology*, 9(2), 257–70.
- Woolfe, T., Want, S. C., & Siegal, M. (2002). Signposts to development: Theory of mind in deaf children. *Child Development*, 73(3), 768–78.
- Yamaguchi, M., Kuhlmeier, V. A., Wynn, K., & VanMarle, K. (2009). Continuity in social cognition from infancy to childhood. *Developmental Science*, 12(5), 746–52.
- Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin*, 124(3), 283–307.
- Youngblade, L. M., & Dunn, J. (1995). Individual differences in young children's pretend play with mother and sibling: Links to relationships and understanding of other people's feelings and beliefs. *Child Development*, 66, 1472–92.

Can theory of mind grow up? Mindreading in adults, and its implications for the development and neuroscience of mindreading

Ian Apperly

Introduction

Why would one study theory of mind in adults? This question would seem ridiculous in almost any other domain of cognition. Yet in more than 30 years of exciting research on mindreading, studies of children and non-human animals have had such a strong grip on the theoretical imagination that it may be difficult even to notice that we do not know how adults do it, let alone to appreciate why we might care. To see how anomalous this situation is, just imagine asking the same question in relation to language or reasoning, or cognition of number, space or causality. In these cases, and for almost any other topic in cognition, there is a long history of research in adults that has yielded core bodies of empirical phenomena and cognitive models that aim to account for them. Yet, despite regular claims for the importance of mindreading for important things that adults do—such as everyday social interaction and communication, moral and legal reasoning—little attention has been paid to how mindreading abilities might need to be implemented in order to perform such roles. However, in recent years this situation has begun to change rapidly. In the first part of this chapter I shall survey this growing literature and advance the view that we need to think of adults as having “two systems” for mindreading.

The absence of cognitive models of mindreading in adults also has unattended consequences in the fields where research has been flourishing. Developmental studies, for all the insights they have given, continue to be conducted with little attention to the mature system that development yields. Indeed, there is almost no research beyond 6 or 7 years of age, as if there were nothing more to mindreading than the ability to pass tests for the minimal possession of key mindreading concepts. And neuroscientific studies, for all their impressive convergence on a “mindreading network” of brain regions, have been limited in their ability to identify the functional contribution of different regions, because unlike other topics in cognitive neuroscience, there have been very limited cognitive accounts of the functions that might be performed. Later in the chapter, I shall discuss how the growing literature on the cognitive basis of mindreading in adults offers new perspectives on neuroscientific and developmental research.

When do we mindread?

Before diving into the findings, it is worth reflecting on what work we believe that mindreading actually does in adults. The success of empirical research on mindreading has led to a tendency

in the field to see mindreading everywhere, so that every communicative exchange and every social interaction is often thought to be mediated by cascading inferences about thoughts, desires, knowledge, and intentions. Interestingly, this tendency runs against some early discussions of mindreading, which emphasized that inferences about mental states were likely to be cognitively demanding, and only made when necessary (e.g. Perner, 1991). It is also inconsistent with suggestions that a great deal of co-ordinated communication can be achieved without mindreading inferences (e.g. Breheny, 2006; Pickering & Garrod, 2004). It has resulted in recent accusations of “theory-of-mind-ism” in research on social interaction, and to suggestions that the “theory of mind” paradigm should be abandoned entirely (e.g. Hutto, 2009; Leudar & Costall, 2009). I believe that a sober assessment of this situation requires us to acknowledge two things. On the one hand, we should not assume that mindreading is at work in a given situation just because the situation can be glossed in such terms. For example, when one person holds a door open for another person whose hands are occupied, it is an open question whether the helper infers the helped person’s intention to open the door themselves, and careful work would be necessary to distinguish this from the possibility that the helper acted on the basis of a social script about door-opening. On the other hand, it is undoubtedly true that we do frequently ascribe mental states to each other, and it is important to understand how and when we do so. The main focus of this chapter will concern this latter point.

It is beyond doubt that in everyday activities we regularly represent the mental states of others. We often tell one another what we think, want or know directly, and in order for this to be understood the listener must, of course, represent these mental states. We also routinely appeal to such mental states when we want to explain or justify the actions of ourselves or others (Malle, 2008). Such circumstances may be relatively trivial, as when I tell you of my desire for beer. But they may also be much more serious, as when we evaluate the guilt or innocence of a defendant in a court of law by considering whether their actions were intentional or accidental, and performed in knowledge or ignorance of their consequences. Viewed this way, mindreading clearly has the potential to be as flexible and as complicated as any other problem of reasoning, and has precisely the wrong characteristics for processing in a specialized cognitive module (Apperly, 2010; cf. Fodor, 1983, 2000). To the degree that this is correct, we should expect mindreading to be relatively effortful, drawing on limited resources for memory and executive control.

On the other hand, it is also commonly supposed that mindreading serves a critical role in fast-moving social interaction and competition, enabling us, for example, to work out what a speaker is talking about on the basis of their eye gaze and to execute competitive bluffs and counter-bluffs in sport. Of course, we must be cautious against theory-of-mind-ism, and remember that mindreading may not always be necessary. However, there seem good *prima facie* reasons for supposing that mindreading inferences are, indeed, made in some such circumstances. To this degree, we should expect mindreading to show at least some key characteristics of a modular process (e.g. Fodor, 1983, 2000; Leslie, 2005), to be relatively effortless, and to make few demands on memory or executive control. Otherwise, the demands of mindreading might detract from our ability to perform the main task at hand, such as acting on a speaker’s request or passing the ball to the best person.

What should be clear, however, is that there is a tension between the requirement that mindreading be extremely flexible, on the one hand, and fast and highly efficient on the other. Such characteristics tend not to co-occur in cognitive systems, because the very characteristics that make a cognitive process flexible—such as unrestricted access to the knowledge of the system—are the same characteristics that make cognitive processes slow and effortful. Instead, flexibility and efficiency tend to be traded against one another. This trade-off is reflected in Fodor’s distinction

between “modular” vs. “central” cognitive processes (Fodor, 1983, 2000). This need for a trade-off is why, in domains as diverse as reasoning (Evans, 2003), social cognition (Gilbert, 1998) and number cognition (Feigenson, Dehane & Spelke, 2004) researchers often propose that human adults have two types of cognitive process operating in that domain, which make complementary trade-offs between flexibility and cognitive efficiency. The above examples suggest that there are good reasons for expecting the same thing for mindreading, and this will be my working hypothesis in the following sections (see Apperly, 2010; Apperly & Butterfill, 2009, for a fuller discussion).

How can we study mindreading in adults¹

Research on children is dominated by questions about the nature and origins of our conceptual understanding of mental states (e.g. Baillargeon, Scott & He, 2010; Perner, 1991; Wellman, Cross & Watson 2001). Typical pass/fail tasks designed to test this conceptual understanding, such as false belief tasks (Wimmer & Perner, 1983) or visual perspective-taking tasks (Flavell, Everett, Croft & Flavell, 1981), are of no use for studying adults because nobody really doubts that a typical adult has such basic mindreading concepts. Researchers have taken several approaches to this problem.

One solution to this problem is to test mindreading concepts that are more subtle or complex, where there might plausibly be some variation among adults. For example, there is evidence that older children and adults advance through a series of increasingly sophisticated theories about the origins and nature of knowledge (e.g. Chandler, Boyes & Ball, 1990; Kuhn, 2009; Robinson & Apperly, 1998). However, such studies are limited by the fact that sophisticated concepts are unlikely to be representative of the mindreading that might underpin many of our everyday social interactions. Other work has shown variation in adults’ ability to understand stories about social situations involving white lies, bluffing, sarcasm, irony or faux-pas (Happé, 1994). Understanding such situations surely requires inferences about the mental states of the story characters. However, it is unclear whether it requires concepts that are more “advanced” than those of younger children. Instead, I would suggest that such tests identify variance in adults’ ability to *apply* such concepts in a flexible, context-sensitive manner. This ability is as vital for everyday mindreading as possessing the concepts in the first place, and plausibly has both an extended developmental course and variability in the mature system of different adults.

A second approach to studying mindreading in adults follows a broad tradition that seeks insights into the nature of adults’ reasoning by examining the heuristics and biases that are apparent in their everyday judgements and decisions. Such studies may pose mindreading problems where the “right” answer is somewhat uncertain, such as judging how another person will make a difficult perceptual discrimination, or interpret ambiguous verbal messages (e.g. Epley, Morewedge, & Keysar, 2004). In tasks with a clear “right” answer—such as predicting the incorrect search of someone with a false belief about an object’s location—researchers may ask participants to rate their certainty about their answer (e.g. Birch & Bloom, 2007; see also Mitchell, Robinson, Isaacs & Nye, 1996). Findings from these studies suggest that adults’ judgements about others are prone to biasing interference from their own perspectives—a phenomenon variously labelled “egocentric bias” (Nickerson, 1999), “reality bias” (Mitchell et al., 1996), and “curse of knowledge” (Birch & Bloom, 2007). Such effects may be most apparent when adults are put under time pressure (Epley et al., 2004), or when placed under a concurrent memory load (Lin, Keysar & Epley, 2010).

¹ For most of the current chapter I will be concerned with methods and findings from typical, neurologically intact adults. Several other chapters discuss research on adults using neuropsychological and neuroimaging methods.

These studies yield valuable insights into the cognitive basis of mindreading, by suggesting that unbiased, non-heuristic mindreading may require time and cognitive effort. However, they give limited insights into why this might be the case, and whether all processing steps in mindreading are cognitively effortful, or only some.

A third approach to studying mindreading in adults uses tasks that require simple judgements about beliefs, desires, and visual perspectives that are conceptually similar to those used in studies of young children. Following methods widely adopted in cognitive psychology these tasks enable the measurement of adults' response times across many repeated trials, and so avoid the problem that adults make few errors on such tasks. For example, in one early study of this kind, German and Hehman (2006) presented adults with multiple trials of a belief-desire reasoning task, which showed adults to be slower to make judgements when a character had a false belief, rather than a true belief, and when s/he had a negative, rather than a positive desire. Because these tasks are simple and repetitive, they may lack the subtlety, sophistication, and uncertainty of much everyday mindreading, which is captured by the tasks described above. However, they have two significant advantages. First, they enable much more fine-grained questions to be asked about the component processes of mindreading. For example, it may be possible to ask whether working memory is necessary for the process of inferring a mental state or the process of using that information to guide social interaction, or both. Secondly, they require simple mindreading concepts similar to those required in most developmental and neuroscientific studies, and so may provide a stronger link to studies of these different participant groups than the methods described above.

In the following sections, I combine evidence from each of these approaches to illustrate what we are learning about the complex nature of mindreading in adults. These findings motivate the suggestion that mindreading can be *both* flexible and effortful, *and* inflexible, effortless, and even automatic.

Mindreading as flexible, but effortful thinking

Discussion about the cognitive basis of mindreading has largely consisted in debates between advocates of simulation-theory, theory-theory and modularity-theory (e.g. Davies & Stone, 1995a,b), which can appear somewhat insular when viewed from outside. The broader literature on cognition in adults already has extensive bodies of research on different aspects of "thinking," including formal and practical reasoning (Byrne, 2005; Johnson-Laird, 1983) and online comprehension during conversation and reading (Garnham, 1987; Pickering & Garrod, 2004). The limited contact between this literature and research on mindreading is truly surprising, because it is almost trivially true that information about mental states—what people know, think, intend, etc.—can and does feature in all aspects of reasoning, decision-making, and discourse processing. Put another way, mindreading is not something we tend to do in isolated and disinterested bouts. Rather, it is an activity that is useful mainly by being part of our everyday thinking and comprehension. This literature is therefore an obvious place to look for expectations about how at least some aspects of mindreading will be achieved.

There are, of course, many alternative accounts of reasoning, decision-making and comprehension that differ in important ways. However, common themes are:

1. Such thinking involves the on-line construction of some form of mental model of the situation under consideration.
2. Models can include information explicitly mentioned (e.g. by the speaker, the story, or in the task instructions) and information from inferences beyond the given information.

3. Model construction and maintenance is demanding of limited resources for memory and executive control.
4. Consequently, what information is represented or inferred will depend upon what memory and executive resources are available, and on whether the thinker takes it to be worthwhile or relevant to elaborate the model.

These themes provide a set of expectations about the characteristics of adults' thinking about thoughts, and there is good evidence to suggest that mindreading does indeed fit these expectations in many circumstances.

Many components of mindreading are effortful

The focus of research on the ages at which children first demonstrate critical mental state concepts might lead to the supposition that the later use of such concepts showed little interesting variability. Yet a number of studies now suggest that mindreading problems that are hardest when children first pass developmentally sensitive tasks (such as false belief vs. true belief problems) continue to require the most cognitive effort for older children and adults. As already mentioned, German and Hehman (2006) presented adults with short stories from which they had to infer a character's belief and desire in order to predict their action. German and Hehman (2006) found that adults were slower (and more error-prone²) on trials that required thinking about false beliefs and negative desires, compared with true beliefs and positive desires, which is the same pattern of relative difficulty observed in 3–6-year-old children on developmentally sensitive tasks (e.g. Leslie, German & Polizzi, 2005). This finding clearly suggests that psychologically relevant parameters, such as the valence of belief and desire, influence the effort adults must put in to solving mindreading tasks. However, in common with most mindreading tasks in the developmental literature, the task required adults to infer the character's mental states from the story, to hold this information in mind and to use it in combination with further facts from the story in order to predict the character's action. This leaves it unclear which of these component processes required cognitive effort.

Further studies have gone some way to isolating these distinct components from one another. Apperly, Warren, Andrews, Grant, & Todd (2011) adapted the belief-desire paradigm and obviated the need for participants to infer the character's mental states by stating these directly. Participants read sentences describing which one of two boxes contained some hidden food, which box the character thought contained the food (his belief could be true or false), and whether he wished to find or avoid the food. All participants had to do was hold this information briefly in mind, and then combine it to predict which box the character would open (e.g. if he had a false belief and a desire to avoid the food he would open the box containing the food on the mistaken belief that this box was empty). Although participants no longer had to infer the character's mental states the valence of his belief (true vs. false) and desire (positive vs. negative), nonetheless, influenced their performance. In a further study, Apperly et al. (2008) obviated both the need to infer a character's mental states and the need to predict their action. Participants read sentences describing the colour and location of a hidden ball and a character's belief about this situation, and responded to a probe

² Even when using very simple mindreading tasks, adult participants do show residual errors. For simple tasks these errors clearly do not reflect a lack of the relevant concepts. Usually, they are either random, with no systematic condition differences, or they follow the same pattern of variation across conditions as are observed in response times. In what follows, I will not mention errors unless they show something interestingly different from response times.

picture that simply required them to recall either belief or reality. Again, false belief trials were harder for participants than a baseline “neutral” belief trial, suggesting that the mere fact of having to hold someone’s false belief briefly in mind comes at a measurable processing cost.

Using a rather different paradigm in which participants made rapid judgements about the simple visual perspective of a character standing in a room, Samson, Apperly, Braithwaite, Andrews, & Bodley (2010) were able to study the demands of mindreading inferences independent of demands associated with withholding such information in mind or using it for further inferences. They found that participants were slower to judge the character’s perspective when it was different from the participants’, suggesting that, like young children, adults experienced egocentric interference when they made judgements about someone else’s perspective. Complementary evidence comes from Keysar, Barr, Balin, & Brauner (2000), who were able to study participants’ ability to *use* information about someone else’s perspective under conditions designed to minimize the demands of inferring this information or holding it in mind. These authors examined adults’ ability to take account of someone’s visual perspective when following their instructions to move objects around a simple array. The instructor could not see all of the items in the array, and so participants had to rule out these items as potential referents for instructions. Importantly, since participants were given ample time to identify these items, and since the array was in full view throughout the trial, any failure to take account of the instructor’s perspective should not be due to difficulty with inferring that perspective or holding that information in mind for an extended period of time. Rather the potential difficulty in this task is with *using* the information about the instructor’s perspective to guide interpretation. In fact, adults are surprisingly error-prone on this task, and indeed they are more error-prone at using the instructor’s perspective than at a comparison condition in which they must interpret instructions according to an arbitrary, non-social rule (Apperly, Carroll, Samson, Qureshi, Humphreys, & Moffatt, 2010).

In summary, recent work shows that it is possible to separate component processes in mindreading—including inferring mental states, holding this information in mind, and using this information. The evidence suggests that these processes may each contribute to making mindreading cognitively effortful. In the next section I review evidence suggesting that much of the variation in “effort” across mindreading problems reflects the differential recruitment of cognitive resources for memory and executive function.

Mindreading frequently depends on memory and executive function

The broader literature suggests that adults’ success on reasoning and comprehension tasks is frequently correlated with their success on tests of memory and executive function, that success is impaired if participants must simultaneously perform a second task that taxes memory or executive function, and that it may also be impaired in old age (e.g. McKinnon & Moscovitch, 2007). A growing literature suggests that the same pattern is typically true for mindreading.

By using a pre-test to select adults with low vs. high working memory spans, Linn, Keysar & Epley (2010) found that adult participants with lower spans were less likely to use their mindreading abilities when following instructions from a speaker with a different visual perspective. By looking for between-task correlations German and Hehman (2006) found that adults’ performance on their belief-desire task was related to performance on tests of inhibitory control, processing speed and working memory, with the most important factors being inhibitory control and processing speed. This study also found that elderly participants (over the age of 60) performed less well than young participants at belief-desire reasoning. Similarly, Phillips, Bull, Allen, Insch, Burr, & Ogg (2011) found that elderly adults performed less well than young adults on false belief tasks (though not true belief tasks), and that this difference was partially mediated by group

differences in working memory performance (see also Mckinnon & Moscovitch, 2007, for similar results). Other studies using tasks that require more subtle or complex mindreading have revealed inconsistent evidence of group differences between younger and older participants, and inconsistent evidence of relationships with other cognitive abilities. However, for the reasons discussed earlier, the demands on mindreading made by more complex tasks are confounded with a range of other requirements on memory, executive function, and context-sensitive processes. This complexity may explain the inconsistent pattern of results observed (see e.g. Rakoczy, Harder-Kasten, & Sturm, 2012 for a recent summary and discussion).

Dual task methods can go beyond correlational studies to provide evidence that concurrent performance of a memory or executive task impairs performance on a mindreading task. This approach has found evidence that mindreading can be impaired by a concurrent working memory task (McKinnon & Moscovitch, 2007), as well as by tasks that tax inhibition and task switching (Bull, Phillips, & Conway, 2008) and verbal repetition (Newton & de Villiers, 2007). However, although these studies show impaired mindreading performance, the tasks used do not make it possible to discern whether participants' difficulty was with mindreading inferences, holding such information in mind or using the information to make inferences about behavior. Two recent studies make some progress on this question. Linn, et al. (2010) found that adults placed under memory load were less able to use information about a speaker's perspective when following their instructions. Qureshi, Apperly and Samson (2010) found that a concurrent inhibitory control task increased adults' egocentric interference when judging another's visual perspective.

In summary, although there is some variation across studies, and some uncertainty about the precise relationships revealed, these studies converge with the evidence from patients with brain injury (see Chapter 10) on the conclusion that mindreading often requires memory and executive function.

Mindreading inferences are non-automatic, and sensitive to context and motivation

It is sometimes stated, simply as a matter of fact, that mindreading inferences are "automatic," suggesting that we cannot help but ascribe mental states when given a stimulus that affords such inferences (e.g. Friedman & Leslie, 2004; Sperber & Wilson, 2002; Stone, Baron-Cohen, & Knight, 1998). Yet, from the perspective of the broader literature on adults' thinking, this claim is surprising. Although there is plenty of evidence that adults routinely and rapidly make inferences that go beyond the information given in a reasoning or comprehension task, it is equally clear that these inferences are not obligatory or stimulus-driven, but are instead dependent on participants' motivation for devoting cognitive resources to this aspect of the task (e.g. Sanford & Garrod, 1998; McKoon & Ratcliff, 1998; Zwaan & Radvansky, 1998). Only recently has evidence begun to bear on this question in relation to mindreading.

Apperly et al. (2006a) presented participants with video scenarios involving a target character who came to have either a true or a false belief about the location of a hidden object. These stimuli clearly afforded mindreading inferences about the character's beliefs, but the instructions only required participants to keep track of the location of the hidden object. Our interest was in whether participants would automatically track the character's belief even though they had no specific reason for doing so. Critical data came from probe questions presented at unexpected points in the videos, which showed participants to be relatively fast at answering questions about the location of the hidden object (which they were instructed to track) but significantly slower to answer matched questions about the character's false belief (which they had not been instructed to track). No such difference in response times to belief and reality probes was found in a second

condition in which participants were instructed to track the character's belief, suggesting that the difference observed in the first condition arose because participants had not inferred the character's belief automatically. Importantly, this finding does not imply that adults only infer beliefs under instruction! Two further studies indicate that varying the scenarios or the context can lead participants to infer beliefs spontaneously (Back & Apperly, 2010; Cohen & German, 2009), and this is a good thing, since the real world does not typically furnish us with explicit prompts to mindread. However, evidence of spontaneous inferences should be distinguished from the claim that mindreading inferences are made in an automatic, stimulus-driven manner, because if inferences are spontaneous then this opens up questions about the contextual conditions that determine the frequency and nature of mindreading.

Important insight into the potential for mindreading to be influenced by contextual factors comes from a study by Converse, Lin, Keysar, & Epley (2008). These authors administered a pre-test in which participants were induced to be in either a happy or a sad mood, and then tested participant's vulnerability to egocentric interference from their own perspective in two different ToM paradigms. Consistent with the view that happy people rely on more heuristic processing, whereas sad people undertake more deliberate processing, these authors found that happy participants showed significantly greater egocentric biases than sad participants. This study not only suggests that mindreading is non-automatic, but that researchers should pay much more attention to the factors that influence the propensity for mindreading, including characteristics of the participant (such as mood) and characteristics of the target, such as their race, sex or class, or other dimensions of similarity and difference to the participant.

It is also important to recognize that the proposition that mindreading inferences are not strictly automatic does not entail that they are typically very slow and effortful. A number of studies arising out of the psycholinguistic tradition suggest that this need not be the case. For example, although there is robust evidence that listeners may fail to take account of the simple perspective of a speaker when interpreting what they say (e.g. Keysar et al. 2000), participants are less likely to look at objects that cannot be seen from the speaker's perspective (e.g. Nadig & Sedivy, 2002), suggesting that information about the speaker's perspective has some cognitive effects (see also Ferguson & Breheny, 2012, for related findings regarding false beliefs). This has led to the suggestion that participants' errors might arise from difficulty with integration of information about the speaker's perspective with linguistic processing of their message (Barr, 2008). However, recent evidence suggests that even such integration of another's perspective need not be very effortful or time-consuming, particularly when no compelling alternative interpretation is available from one's own perspective (Ferguson & Breheny, 2011).

Altogether, there is direct evidence to suggest that mindreading frequently occurs spontaneously. In a wide range of circumstances people clearly do not need to be explicitly directed to take account of what other people see, think or feel. Nonetheless, these inferences are not automatic, and the likelihood of spontaneous mindreading depends on the context and on the participant's mood. The broader literature on inferences made during discourse provides compelling grounds for thinking that future work will find that the likelihood of spontaneous mindreading, as well as the extent of elaboration of such inferences, will depend on participants' motivation and on the availability of cognitive resources for memory and executive control.

Summary

The view of mindreading that emerges from research reviewed in the sections above is as follows. At one extreme end of the scale, exemplified by a jury's deliberations about the evidence for and against a defendant having acted knowledgeably and intentionally, mindreading may be truly slow,

deliberative, and effortful. However, the general literature concerning inferences made online during comprehension should lead us to expect that many mindreading inferences may often be made without too much deliberative scratching of chins, and used quickly enough to keep up with an unfolding discourse or text. Nonetheless, such mindreading will require cognitive effort and will depend on the availability of the necessary motivation and cognitive resources.

Mindreading as a cognitively efficient, but inflexible and limited process

Discussions about the possible automaticity of mindreading typically underestimate how difficult it is to determine that a cognitive process is performed in an automatic manner (e.g. Moors & De Houwter, 2008). For example, it is certainly not sufficient to show that mindreading occurs without instruction, or even that it occurs relatively quickly. For as already described, much cognitive processing can occur spontaneously and quite rapidly, but the fact that it does so only when participants are appropriately motivated and have sufficient resources suggests that such processing is not automatic. However, there are good reasons in principle for thinking that at least some mindreading needs to be less like “thinking” and more like perception in character. To this degree, we should expect mindreading processes to be less dependent on participants’ motivations or cognitive resources, and also to be more limited in their scope than the ones described so far. Recent research also lends support to this view of mindreading.

Evidence that mindreading may occur when unnecessary or unhelpful

One characteristic of processes that are more perception-like or modular is that they occur at least somewhat independently of participants’ motivation or purpose, and may even interfere with their primary objectives. Evidence from three different paradigms suggests that mindreading may sometimes show such characteristics.

Zwicker (2009) presented participants with very simple animations of isosceles triangles that appeared to be moving in a random fashion, in a simple goal-directed fashion (e.g. one triangle chased another), or in a complex goal-directed fashion (e.g. one triangle coaxed another). Previous research has found that participants’ spontaneous descriptions of these animations differ, with simple and complex goal-directed animations eliciting descriptions of goals, and only complex goal-directed animations eliciting descriptions of more complex mental states (Abell, Happé, & Frith, 2000). During the animations a dot occasionally appeared on one or other side of a triangle and participants’ explicit task was to judge whether the dot appeared to the left or the right. On half of the trials the triangle happened to be pointing upwards when the dot appeared, and on the other half it happened to be pointing downwards. Of course, this was strictly irrelevant to the participants’ task of making left-right judgments of the dots. Nonetheless, in the two goal-directed conditions participants were slower to make left-right judgements for downward-facing triangles than for upward-facing triangles, whereas there was no such effect for the random movement condition. It is notable that, if a triangle is perceived to have a “perspective,” then for upward-facing triangles, the triangle’s left or right was aligned with the participant’s own left and right, whereas the left side of a downward-facing triangle was on the participants’ right side, and vice versa. Thus, participants’ slow left-right judgments for downward-facing goal-directed triangles can be understood as being the result of interference from task-irrelevant processing of the triangle’s “perspective” in the goal-directed conditions but not the random condition. It is notable that this effect was largest of all for the complex goal-directed animations. However, it is not clear whether this was because these stimuli invited the richest ascriptions of mental states, or because these stimuli gave

the more compelling sense of animacy. Nor is it clear whether participants were processing the triangle's physical, spatial perspective, or whether they were, in some sense, attributing a psychological, visual perspective to the triangle. Either would be sufficient to support a left-right distinction. Importantly, though, this does appear to be a case in which some form of perspective-taking is occurring independently of participants' purposes, and in fact this interferes with their performance on the main task.

A second paradigm converges on the same conclusions, this time in the case of very simple visual perspective-taking. Samson et al. (2010) presented participants with pictures of a room with dots on the wall, and an avatar positioned in the room such that he either saw all of the dots (so his perspective was congruent with participants') or he saw a subset of the dots (so his perspective was incongruent with participants'). On the trials that are critical for the current discussion, the avatar's perspective was irrelevant because participants were simply asked to judge how many dots they saw in the room from their own "self" perspective. Nonetheless, participants' responses were slower when the avatar's perspective happened to be incongruent, rather than congruent with their own. This effect was apparent when participants' "self" judgements were mixed with other trials on which they made explicit judgments about the avatar, and also in a further experiment in which participants only ever made judgements about their own perspective. In the latter case, the avatar's perspective was entirely irrelevant to the entire task, and yet participants appeared to process his perspective, and this caused interference when it differed from their own.

A third paradigm converges on related conclusions, this time for processing of belief-like states.³ Kovács, Téglás, & Endress (2010) presented participants with animations in which a ball rolled around a scene, sometimes appearing to remain behind an occluding wall, and sometimes rolling out of the scene. The animations also included an agent who witnessed different parts of the event sequence across trials and ended up either with the same belief as the participant about the ball's presence or absence, or the opposite belief. However, the agent was irrelevant to the participants' task, because participants were simply required to press a response button if the ball was behind the wall when the wall was lowered at the end of the animation. The ball was, in fact, equally likely to be present irrespective of whether it had appeared to remain or to leave the scene during the animation. Unsurprisingly, adults were faster to detect the ball when the animation led them to expect the ball to be present than when they expected it to be absent. Importantly, though, this effect was modulated by the irrelevant beliefs of the agent: when the ball was unexpectedly present from the participants' point of view participants were faster to detect it if the agent happened to believe that it was present and slower to detect it when the agent happened to believe it was absent. In this case, processing of the agent's perspective was actually helpful, rather than unhelpful, but nonetheless it was clearly irrelevant to participants' main task of detecting balls appearing behind the wall, suggesting that it was relatively stimulus-driven and automatic.

Evidence that mindreading is cognitively efficient

A second characteristic of perception-like, modular processing is that it makes few demands on domain-general resources for its operation. One way to test this experimentally is to see whether effects such as those just described persist even when participants' resources are taxed by another task.

³ It is a moot point whether adults or infants in such paradigms are representing beliefs per se, or simpler belief-like states (Apperly & Butterfill, 2009). However, what is critical here is that interference arises in a situation where the agent has a false belief, rather than a different visual perspective.

Qureshi, et al. (2010) presented Samson et al.'s visual perspective-taking task either alone or at the same time as a task that taxed executive control. Their rationale was that participants' irrelevant processing of the avatar's perspective might, nonetheless, be consuming of executive resources, and if this were so then the secondary task should reduce this irrelevant processing and so reduce the interference that participants suffered when judging their own perspective. In fact, this study found that the secondary task *increased* interference from the avatar's irrelevant perspective, suggesting that calculating his perspective was cognitively efficient, and that executive control was instead required for resisting interference from this perspective.

Evidence for the same conclusion comes from Schneider, Bayliss, Becker, & Dux (2012). These authors monitored adults' eye fixations while viewing video scenarios in which the character in the video came to have either a true or a false belief about an object's location. Although adults always knew the object's true location and although the character's beliefs were apparently irrelevant, adults nonetheless spent longer looking at the incorrect location for the object when this was where the character incorrectly believed the object was located, compared with when the character had a true belief. Importantly, adults showed no awareness of tracking the character's beliefs, and this evidence of "implicit" processing was replicated in a second study in which participants simultaneously performed a distracting secondary task.

The findings from these two studies suggest that simple visual perspective-taking and simple belief ascription not only occur in a relatively automatic manner, but also can be cognitively efficient so that these processes are not disrupted by a secondary task.

Evidence that mindreading is limited

A third characteristic of perception-like, modular processing is that automaticity and efficiency do not come for free, but are gained at the expense of limits on the kinds of problem that can be solved. A well-studied example is the ability of infants, children, adults, and many non-human species to track the precise numerosity of items in a set (see e.g. Feigenson et al., 2004). This ability is cognitively efficient, but also extremely limited, in that it can only "count" to 3. Importantly, such limitations are not merely a correlate of modular processing; limits reflect the way in which modular processing manages to be efficient, by restricting itself to processing of just some kinds of information (e.g. Fodor, 1983, 2000). It follows that, to the degree that mindreading shows other characteristics of modular processing, we should expect it also to be limited to process some problems, but not others.

A recent study that fits with this expectation of limited processing was conducted by Surtees, Butterfill and Apperly (2012; see also Low & Watts, in press, described later). These authors tested whether Samson et al.'s (2010) finding that adults automatically process *what* items were seen by an avatar in a cartoon room would extend to *how* items were seen by the avatar. In their task the avatar faced out of the room, sitting behind a table on which digits could appear. Digits such as the number "8" are rotationally symmetrical, and so would appear the same to both the avatar and the participant. These trials were compared with others using digits such as the number "6" that would look like a "six" to one viewer and a "nine" to the other. Recall that Samson et al. (2010) found that participants were slower to judge *what* they themselves could see when the avatar saw something different. In contrast, Surtees et al. (2012) found no evidence that adults were slower to judge *how* the digit appeared to them when it happened to appear differently for the avatar. Naturally, we must be cautious about drawing strong conclusions from these negative findings, but nonetheless this study provides preliminary evidence fitting with the expectation that automatic mindreading will be limited in its scope.

Interim summary: two systems for mindreading in adults?

The foregoing sections show that recent research has greatly extended the methods available for studying mindreading in adults. However, on key questions about the cognitive characteristics of mindreading the results emerging from this work point in quite different directions. Some evidence suggests that mindreading shows the characteristics of flexible, but effortful thinking, while other evidence suggests that it shows the characteristics of efficient, but inflexible modular processing. What are we to make of these findings? There is certainly some wisdom in the view that we should be cautious. Many of the paradigms described are novel, at least within the mindreading literature, and most findings reported are relatively new. Any new field of enquiry is likely to produce a higher than average number of anomalous findings, and in 5 or 10 years there might be a much better evidence base to suggest that mindreading is more like thinking than perception, or vice versa.

However, I think there are good grounds for taking both characterizations of mindreading seriously. First, the findings from adults may be relatively new, but the evidence comes from multiple tasks and approaches that provide reassuring convergence suggesting that both characterizations of mindreading have merit. Secondly, the evidence base is potentially much broader if we also look to studies of children and infants. Here, too, we find good grounds for supposing that mindreading has the characteristics of effortful thinking when studied in children (e.g. Carlson & Moses, 2001), but also apparently contradictory evidence that it has more perception-like qualities when studied in infants (Baillargeon, et al., 2010). Thirdly, such apparently contradictory results abound in psychological research in other cognitive domains, such as number and physical cognition, social cognition and general reasoning (e.g. Evans, 2003; Feigenson et al., 2004; Gilbert, 1998). In these other domains, this apparent contradiction is resolved by supposing that adults actually operate with “two systems,” each having distinct processing characteristics. For these reasons, it seems at least plausible to hypothesize that adults implement two kinds of solutions for mindreading, consisting both of flexible processes for “thinking” about the minds of others, and a number of modules that pull off the same trick in a cognitively efficient manner for a limited subset of mindreading problems (e.g. Apperly & Butterfill, 2009; Apperly, 2010).

Understanding the cognitive basis of mindreading in adults is surely a worthwhile project in its own right. However, it also has further utility in informing our understanding of development and neural basis of mindreading. In the final sections of this chapter I shall explore some important implications of the emerging evidence about the multi-faceted nature of mindreading in adults.

Implications for development

The growing literature on mindreading in adults should have a significant impact on studies of development for several reasons. First, it is producing new methods based upon the measurement of response times that can be adapted for use with “older” children who pass standard developmental tests of mindreading. Such methods suggest that children’s use of information about the minds of others becomes significantly more accurate through middle childhood and adolescence (Dumontheil, Apperly, & Blakemore, 2010; Epley et al., 2004), that different belief-desire reasoning problems continue to vary in difficulty even after children first “pass” the tasks (Apperly et al., 2011), and that 6-year-olds show just the same degree of automatic perspective-taking as adults (Surtees & Apperly, 2012). Secondly, the cognitive basis of mindreading in adults can assist with interpretation of developmental findings. For example, there is good evidence that adults who have severely impaired grammar as a result of brain injury may nonetheless be able to pass both 1st and 2nd order mindreading tasks (e.g. Apperly et al., 2006b; Varley & Siegal, 2000;

Varley, Siegal & Want, 2001). This suggests that developmental associations between grammar and mindreading (e.g. Milligan, Astington & Dack, 2007) cannot be the result of grammar having a constitutive role in the mature mindreading system that children are developing, but must instead be due to grammar serving a role in the developmental construction of mindreading (Apperly, Samson & Humphreys, 2009). Such conclusions are difficult to reach without evidence from adults. Thirdly, the adult system is the end-point that any adequate theory of development must be able to explain. It should be clear from the complex picture of the adult system described above that developmental accounts focusing only on when infants or children should be credited with basic mindreading concepts are in danger of seriously underestimating their explanatory task. This is so because such accounts often have rather little to say about what happens after children pass basic experimental paradigms. In the following paragraphs I will consider in very broad terms what questions a two-systems account of mindreading in adults should make us ask about development.

How do adults acquire two systems for mindreading?

As described in other chapters of this volume, most research on mindreading in children focuses on 2–6-year-olds and suggests that the ability to make correct judgements about other people's beliefs, desires and intentions has a protracted developmental course. Not only is there good evidence of incremental acquisition of an increasingly sophisticated conceptual grasp of mental states (Wellman & Liu, 2004), but progress appears to depend critically on developments in both language, and executive function and memory (e.g. Carlson & Moses, 2001; Milligan, Astington & Dack, 2007). Although such research seldom looks much beyond early childhood, it seems natural to see this as charting the early development of the adult system for mindreading that has the characteristics of flexible, but cognitively effortful thinking.

Because of this well-known body of findings in children, much excitement has attended recent evidence suggesting that infants are also capable of mindreading, at least when tested using methods that allow this ability to be observed in eye movements, looking time or other spontaneous behaviors, rather than in overt judgements (see e.g. Baillargeon et al., 2010 for a recent review). Much of the excitement concerns the simple possibility that mindreading might be observed at much younger ages than previously thought. However, just as interesting from a cognitive point of view is the fact that infants' mindreading must be cognitively efficient, since infants have few resources for language or executive control. The findings from infants remain somewhat controversial (e.g. Hutto, Herschbach & Southgate, 2011; Perner, 2010), but for current purposes I shall work with the hypothesis that infants are indeed mindreading in some meaningful sense. Instead, the question on which I would like to focus is how the abilities of infants are related to those of older children and adults.

The infant system grows up

The dominant view among researchers studying mindreading in infants appears to be that infants' abilities will be essentially continuous with the full-blown mindreading abilities of older children and adults (e.g. Baillargeon et al., 2010; Leslie, 2005). That is to say, infants possess foundational mindreading concepts and abilities that are, at first, only "implicit" and only observable via indirect experimental methods. However, over developmental time, and with increasing availability of language, executive function and critical social knowledge, children become increasingly able to use these concepts in flexible and sophisticated ways, and to use them as the basis for explicit judgements. This developmental pattern is depicted in the top panel of Figure 5.1, and is clearly

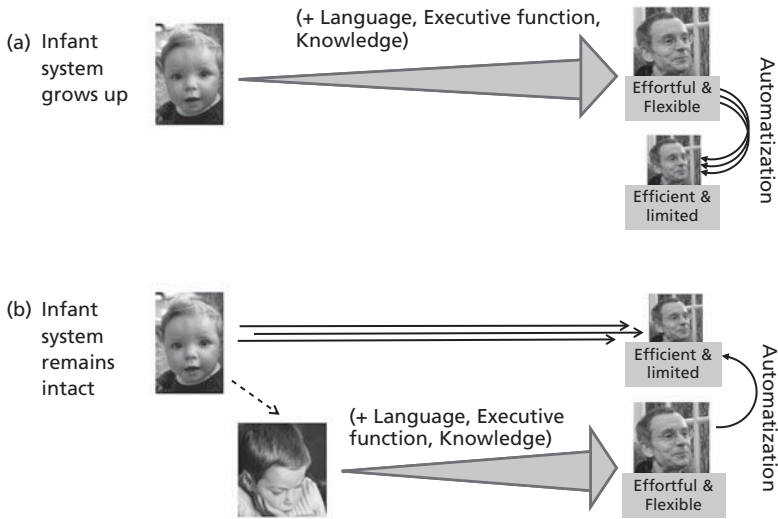


Figure 5.1 Alternative relationships between the mindreading abilities of infants, children and adults. In panel (a) the mindreading abilities infants become progressively integrated with language, executive function and knowledge over the course of development, giving rise to the flexible but effortful abilities of adults. Some of these abilities are then automatized into efficient, but inflexible routines. In panel (b) the mindreading abilities of infants remain largely intact into adulthood, where they enable adults to perform some mindreading in an efficient but inflexible manner. Young children undergo a protracted process of learning to reason about the minds of others. This developmental process requires language, executive function, and accumulating knowledge, and gives rise to the flexible, but effortful mindreading abilities of adults. The dashed lines suggest that this model is compatible with the abilities of infants having some influence on the development of children's reasoning about mental states, and with adults automatizing some of their effortful mindreading abilities.

NB. I am grateful to Oliver Poole and Lionel Apperly for allowing their photographs to be used.

plausible as an account of the relationship between the abilities of infants and adults. However, one important consequence of this proposal is that although infants' abilities may start out being cognitively efficient, they will clearly not remain so once they have been integrated with language, executive function and an ever-increasing database of knowledge. This follows from Fodor's (1983, 2000) analysis, which holds that it is precisely the absence of such integration that explains how modular processing can be cognitively efficient.

This means that there must be some additional developmental explanation of the cognitively efficient mindreading abilities of adults. One potential way in which this might occur is that certain mindreading problems that are both sufficiently frequent and sufficiently regular in their demands will become automatized into routines. For example, it might be that over developmental time most people encounter the need to calculate what someone sees with sufficient frequency that this becomes automatized, so that "what someone sees" is calculated whenever we see an agent apparently attending to objects in her visual field.

The infant system remains intact

Importantly, though, this is not the only possible set of developmental relationships. As depicted in the bottom panel of Figure 5.1, another possibility is that the abilities observed in infants remain

intact and uncluttered by demands upon language or executive control, so that they continue to support cognitively efficient mindreading into adulthood. On this account, although the infant system may provide critical support, flexible and effortful mindreading would develop as a quite separate process, perhaps in much the way envisaged by developmental psychologists before the recent findings from infants. Of course, this hypothesis does not preclude the possibility that some initially effortful mindreading might become automatized over developmental time, but only on this hypothesis will adults inherit at least some of their efficient capacities for mindreading from infants.

Clearly, these accounts present quite different views of the developmental origins of adults' two systems for mindreading. So how might we decide between them? Although the most popular current suggestion is that the infant system grows up, it is noteworthy that many other domains of cognition, such as number, physical cognition, agency and causality, there is good evidence for the alternative account, that infant abilities remain intact into adulthood (for a recent extensive review and discussion, see Carey, 2009). Of course, these precedents alone are insufficient to show that the same will be true for mindreading. However, Carey's account does indicate where decisive evidence might be found; in the nature of the limits on efficient mindreading observed in infants and adults. Recall from earlier that efficient mindreading in both infants and adults will necessarily come at the cost of inflexible limits on the kinds of information that can be processed. In adults this will be the case irrespective of whether efficient mindreading abilities are inherited from infants, or whether they are automatized. However, if adults' efficient abilities arise as a result of automatization, then there is no reason to suppose that these limits will be the same as those observed in infants. If, on the other hand, adults inherit efficient mindreading abilities from infants, then they should show similar limits. This is the case for number cognition, where both infants' and adults' capacities for precise enumeration are limited to three items. As Carey (2009) points out, such "signature limits" are a powerful device for detecting whether infants and adults are using the same cognitive processes to solve a problem. Stephen Butterfill and I have argued elsewhere that there is indeed preliminary evidence for such a signature limit in the abilities of infants and the efficient abilities of adults, which both may be restricted to process relations between agents and objects, rather than agents and propositions (Apperly, 2010; Apperly & Butterfill, 2009; Butterfill and Apperly, *in press*). And this proposal has received recent support from Low and Watts (*in press*) who find evidence that young children's "implicit" understanding of false belief allows them to ascribe false beliefs about an object's location but not about an object's identity. However, this specific proposal matters less in the current context than the general proposition that there is more than one way in which the mindreading abilities of infants can develop through childhood into those we observe in adults, and that there are viable ways of distinguishing between these developmental hypotheses.

Implications for understanding the neural basis of mindreading

Research on the neural basis of mindreading has been strongly influenced by traditional developmental approaches, with two notable consequences. First, the tasks employed typically involve presentation of stories or cartoons that resemble tests of young children's explicit reasoning about mental states, and so these tasks should be expected to test adults' "thinking" about mental states. Secondly, studies are typically premised on the assumption that mindreading consists primarily in the domain-specific ability to understand and represent mental states (e.g. Frith & Frith, 2003; Saxe, Carey & Kanwisher, 2004). Thus, although a number of brain areas are commonly held to constitute a "mindreading" brain network, notably including medial prefrontal cortex (mPFC), temporal poles and bilateral temporo-parietal junction (TPJ), debate has typically been limited

to asking which of these areas is most selectively involved, and might therefore qualify as the location of the neural seat of mindreading. Perhaps the clearest evidence emerging from this line of thinking comes from a series of studies by Saxe and colleagues (e.g. Saxe & Kanwisher, 2003; Saxe & Powell, 2006). These authors first identified brain areas that survived a very neat contrast between activation observed while participants responded to short stories concerning false beliefs vs. false photographs, and then tested which of these areas were most selectively activated during other judgements about mental states in contrast to other judgements, including personal preferences, personality, physical appearance. These studies consistently find that right-TPJ shows the largest and most selective activation for mental states, whereas other areas of the “mindreading network” either show lower activations, or activations for a wider range of judgements. This pattern has led to suggestions that r-TPJ is *the* domain-specific neural basis of mindreading (e.g. Saxe, 2006).

This interpretation of these studies remains highly contested (e.g. Decety & Lamm, 2007; Mitchell, 2008; Legrand & Ruby, 2009), but it is not my current interest to enter into this debate. Studies of the cognitive basis of mindreading, reviewed above, clearly do nothing to rule in or rule out the possibility that there are genuinely domain-specific representations and processes involved in mindreading. However, they do suggest very clearly that there is a great deal more to mindreading than possessing specialized mindreading concepts or representations. At the very least, doing useful work with such concepts will involve the ability to make flexible inferences in a context-sensitive manner, to do this within the context of a mental model of the on-going situation, and all the while to resist interference from self-perspective. These considerations suggest that the benefits of tightly-controlled subtractive methods for identifying neural activation that could be specific to mindreading will likely come at a cost. In particular, they risk causing researchers to overlook functional and neural processes that are less specific, but equally essential to a full understanding of mindreading. Therefore I will briefly focus on studies using different methods, which cast light on the broader neural basis of mindreading.

Medial prefrontal cortex features prominently among the other neural regions implicated in mindreading. However, in the broader literature mPFC is also implicated in a range of other tasks, including generation of temporary integrated representations of events, and imposing structure on otherwise vague or uncertain problems (see e.g. Legrand & Ruby, 2009). As discussed above, mindreading frequently requires context-sensitive inferences, made on the fly, using limited information about the situation. Might it be, then, that mPFC is involved in serving this role for mindreading? In a recent study, Jenkins and Mitchell (2009) presented participants with mindreading tasks that orthogonally varied whether the scenarios concerned a character’s mental states or their preferences, and whether a specific mindreading inference was relatively clear, given the context, or whether the situation was more ambiguous. Consistent with other work, this study found that r-TPJ was selectively sensitive to the difference between scenarios involving mental states rather than preferences, whereas mPFC was not selectively sensitive to this difference. In contrast, mPFC was sensitive to the difference between scenarios involving clearly-specified rather than ambiguous inferences, whereas r-TPJ was not. Naturally, this finding must be interpreted with some caution given how little agreement there is on the functions of mPFC in general (see e.g. Legrand & Ruby, 2009), or the functional necessity of mPFC for mindreading in particular (e.g. Bird Castelli, Malik, Frith, & Husain, 2004). Nonetheless, it serves to illustrate how it is possible to go beyond asking which regions of the “mindreading network” are most specifically involved in mindreading, in order to understand how the multiple functional requirements of mindreading are fulfilled.

Not only is it the case that commonly-used subtractive methods bias researchers to ask just one kind of question about regions of the “mindreading network”, but they also risk leading researchers to overlook additional functional and neural processes that might be critically necessary for mindreading. One such illustration comes from the case study of a patient, WBA, who, following a stroke, sustained a right frontal lesion that only showed limited encroachment on regions of the “mindreading network,” but encroached substantially on brain regions frequently implicated in cognitive control (Samson, Apperly, Kathirgamanathan, & Humphreys, 2005). WBA showed impairment on a range of neuropsychological tests for working memory and executive function, including inhibitory control. Across a range of mindreading tasks he showed a pronounced tendency for “egocentrism”, responding on the basis of his own belief, desire or perspective, rather than that of the other person. Nonetheless, on a false belief task designed to reduce the tendency for egocentrism by reducing the salience of participants’ self-perspective, WBA was able to perform successfully. These results indicate that having the ability in principle to think about someone else’s perspective is not nearly sufficient for reliable mindreading. To put that ability into practice in a typical range of circumstances also requires the ability to inhibit interference from one’s own perspective, and this ability was impaired by WBA’s right frontal lesion. This conclusion receives converging support from several functional magnetic resonance imaging (fMRI) and event-related potential (ERP) studies using designs that that manipulate demands on self-perspective inhibition within the context of a mindreading task (e.g. McCleery, Surtees, Graham, Richards, & Apperly, 2011; van der Meer, Groenewold, Nolen, Pijnenborg, & Aleman, 2011; Vogeley et al., 2001). These studies show lateral frontal brain regions—notably inferior frontal gyrus—being recruited in the service of mindreading. Such activation is not observed in the most tightly-controlled subtraction designs—such as the comparison between false belief and false photograph tasks—because the relevant activation is subtracted out.

What I hope this brief section illustrates is that emerging evidence on the cognitive basis of mindreading in adults has significant consequences for how neuroscientific investigations of mindreading are designed and interpreted. The large number of studies that seek to identify the neural basis of domain-specific mindreading processes make a valuable contribution to understanding. However, there are strong grounds for thinking that this will be just one part of a full account of the neural basis of mindreading.

General conclusion

For more than 30 years research on our ability to understand agents in terms of mental states has been remarkably productive, but at the same time surprisingly narrow in its scope. We have learned a great deal about how and when children first come to mindread, the degree to which these abilities are shared with other species, and, most recently, the neural basis of some aspects of mindreading. However, we have only scratched the surface of understanding the mature abilities that children develop, and how adults use these abilities on-line as they communicate and socialize, or talk, read, and think about mental states. This situation is changing rapidly, and it is motivating changes in how we conceptualize mindreading. In addition to answering questions about who has mindreading concepts and when they have them, an adequate theory of mindreading must explain how we ever make use of such abilities. In particular, it must explain how we manage to be both extremely subtle and sophisticated mindreaders, yet simultaneously achieve at least some mindreading rapidly enough to keep up with fast-moving social interactions. I hope to have made the case that mindreading in adults is not merely a fast-emerging new sub-topic in the mindreading literature, but that it is providing critical new insights about the nature of mindreading itself.

References

- Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, 15(1), 1–16.
- Apperly, I. A. (2010). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Hove: Psychology Press/Abingdon: Taylor & Francis Group.
- Apperly, I. A. & Butterfill, S. A., (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4), 953–70.
- Apperly, I. A., Carroll, D. J., Samson, D., Qureshi, A., Humphreys, G. W. & Moffatt, G. (2010). Why are there limits on theory of mind use? Evidence from adults' ability to follow instructions from an ignorant speaker. *Quarterly Journal of Experimental Psychology*, 63(6), 1201–17.
- Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adult performance on a non-inferential theory of mind task. *Cognition*, 106, 1093–108.
- Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C. & Samson, D. (2006a). Is belief reasoning automatic? *Psychological Science*, 17(10), 841–4.
- Apperly, I. A., Samson, D., & Humphreys, G. W. (2009). Studies of adults can inform accounts of theory of mind development. *Developmental Psychology*, 45(1), 190–201.
- Apperly, I. A., Samson, D., Carroll, N., Hussain, S., & Humphreys, G. W. (2006b). Intact 1st and 2nd order false belief reasoning in a patient with severely impaired grammar. *Social Neuroscience*, 1(3–4), 334–48 (Special issue on theory of mind).
- Apperly, I. A., Warren, F., Andrews, B. J., Grant, J. & Todd, S. (2011). Error patterns in the belief-desire reasoning of 3- to 5-year-olds recur in reaction times from 6 years to adulthood: evidence for developmental continuity in theory of mind. *Child Development*, 82(5), 1691–703.
- Back, E., & Apperly, I. A. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition*, 115(1), 54–70.
- Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14, 110–18.
- Barr, D. J. (2008) Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, 109(1), 18–40.
- Birch, S. A. J. & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, 18(5), 382–6.
- Bird, C. M., Castelli, F., Malik, O., Frith, U., Husain, M. (2004) The impact of extensive medial frontal lobe damage on 'theory of mind' and cognition. *Brain*, 127(4), 914–28.
- Breheny, R. (2006) Communication and folk psychology, *Mind & Language*, 21(1), 74–107.
- Bull, R., Phillips, L. H. & Conway, C. (2008). The role of control functions in mentalizing: Dual task studies of theory of mind and executive function. *Cognition*, 107, 663–72.
- Butterfill, S. & Apperly I. A. (In press). How to construct a minimal theory of mind. *Mind & Language*.
- Byrne, R. M. J. (2005). *The Rational Imagination: How People Create Alternatives To Reality*. Cambridge: MIT Press.
- Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, 72, 1032–53.
- Carey, S. (2009) *The Origin of Concepts*. Oxford: Oxford University Press.
- Chandler, M., Boyes, M., & Ball, L. (1990). Relativism and stations of epistemic doubt. *Journal of Experimental Child Psychology*, 50, 370–95.
- Cohen, A. S. & German, T. C. (2009). Encoding of others' beliefs without overt instruction. *Cognition*, 111, 356–63.
- Converse, B. A., Lin, S., Keysar, B., & Epley, N. (2008). In the mood to get over yourself: Mood affects theory-of-mind use. *Emotion*, 8, 725–30.
- Davies, M. & Stone, T. (Eds). (1995a). *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell.

- Davies, M. & Stone, T. (Eds). (1995b). *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell.
- Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *Neuroscientist*, 13, 580–93.
- Dumontheil, I., Apperly, I. A., & Blakemore, S. J. (2010). Online use of mental state inferences continues to develop in late adolescence. *Developmental Science*, 13(2), 331–8.
- Epley, N., Morewedge, C., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, 40, 760–8.
- Evans, J. St. B. T. (2003) In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10), 454–9.
- Feigenson, L., Dehane, S. & Spelke, E. S. (2004) Core systems of number. *Trends in Cognitive Sciences*, 8(7), 307–14.
- Ferguson, H. J., & Breheny, R. (2011). Eye movements reveal the time-course of anticipating behavior based on complex, conflicting desires. *Cognition*, 119, 179–96.
- Ferguson, H. J. and Breheny, R. (2012). Listeners' eyes reveal spontaneous sensitivity to others' perspectives. *Journal of Experimental Social Psychology*. 48, 257–63.
- Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual-perception—further evidence for the level 1-level 2 distinction. *Developmental Psychology*, 17, 99–103.
- Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge: MIT Press.
- Fodor, J. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge: MIT Press.
- Friedman, O., & Leslie, A. M. (2004). Mechanisms of belief-desire reasoning: Inhibition and bias. *Psychological Science*, 15, 547–52.
- Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalising. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 358, 459–73.
- Garnham, A. (1987). *Mental Models as Representations of Discourse and Text*. Chichester: Ellis Horwood.
- German, T. P. & Hehman, J. A. (2006) Representational and executive selection resources in “theory of mind”: Evidence from compromised belief-desire reasoning in old age. *Cognition*, 101, 129–52.
- Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds), *The Handbook of Social Psychology* (4th edn, pp. 89–150). New York: McGraw Hill.
- Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24, 129–54.
- Hutto, D. D. (2009). Folk psychology as narrative practice. *Journal of Consciousness Studies*, 16(6–8), 9–39.
- Hutto, D. D., Herschbach, M. & Southgate, V. (2011) Social cognition: Mindreading and alternatives. Editorial to the special issue. *Review of Philosophy and Psychology*, 2(3), 375–95.
- Jenkins, A. C. & Mitchell, J. P. (2009). Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex*, 21(8), 1560–70.
- Johnson-Laird, P. N. (1983) *Mental Models: Toward a Cognitive Science of Language, Inference, and Consciousness*. Cambridge: Cambridge University Press.
- Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychological Sciences*, 11, 32–8.
- Kovács, Á. M., Téglás, E. & Endress, A. D. (2010). The social sense: Susceptibly to others' beliefs in human infants and adults. *Science*, 330, 1830–4.
- Kuhn, D. (2009). The importance of learning about knowing: Creating a foundation for development of intellectual values. *Child Development Perspectives*, 3(2), 112–17.
- Legrand, D., Ruby P. (2009) What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychological Review*, 116(1), 252–82.

- Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, 9(10), 459–62.
- Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, 50, 45–85.
- Leudar, I & Costall, A. (2009) *Against Theory of Mind*. Basingstoke: Palgrave Macmillan.
- Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46, 551–6.
- Low, J. & Watts, J. (in press). Attributing false-beliefs about object identity is a signature blindspot in humans' efficient mindreading system. *Psychological Science*.
- Malle, B. F. (2008). Fritz Heider's legacy: Celebrated insights, many of them misunderstood. *Social Psychology*, 39, 163–73.
- McCleery, J. P., Surtees, A., Graham, K. A., Richards, J. & Apperly, I. A. (2011). The neural and cognitive time-course of theory of mind. *Journal of Neuroscience*. 31(36): 12849–54.
- McKinnon, M. C. & Moscovitch, M. (2007) Domain-general contributions to social reasoning: Theory of mind and deontic reasoning re-explored. *Cognition*, 102(2), 179–218.
- McKoon, G. & Ratcliff, R. (1998). Memory-based language processing: Psycholinguistic research in the 1990s. *Annual Review of Psychology*, 49, 25–42.
- Milligan, K., Astington, J. W. & Dack, L. A. (2007) Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*. 78(2), 622–46.
- Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, 18(2), 262–71.
- Mitchell, P., Robinson, E. J., Isaacs, J. E. & Nye, R. M. (1996). Contamination in reasoning about false belief: An instance of realist bias in adults but not children. *Cognition*, 59, 1–21.
- Moors, A. & De Houwter, J. (2008). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, 132(2), 297–326.
- Nadig, A. S., & Sedivy, J. C. (2002). Evidence for perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, 13, 329–36.
- Newton, A. M. & de Villiers, J. G. (2007) Thinking while talking: Adults fail nonverbal false belief reasoning. *Psychological Science*, 18 (7), 574–9.
- Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, 125, 737–59.
- Perner, J. (1991). *Understanding the Representational Mind*. Brighton: Harvester.
- Perner, J. (2010). Who took the cog out of cognitive science? – Mentalism in an era of anti-cognitivism. In: P. A. Frensch & R. Schwarzer (Eds), *Cognition and Neuropsychology: International Perspectives on Psychological Science* (Vol. 1, pp. 241–61). Hove, UK: Psychology Press.
- Phillips, L. H., Bull, R., Allen, R., Inch, P., Burr, K. & Ogg, W. (2011). Lifespan aging and belief reasoning: Influences of executive functions and social cue detection. *Cognition*, 120, 236–47.
- Pickering, M. J. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169–226.
- Qureshi, A., Apperly, I. A. & Samson, D. (2010). Executive function is necessary for perspective-selection, not Level-1 visual perspective-calculation: Evidence from a dual-task study of adults. *Cognition*, 117(2), 230–6.
- Rakoczy, H., Harder-Kasten, A., & Sturm, L. (2012). The decline of theory of mind in old age is (partly) mediated by developmental changes in domain-general abilities. *British Journal of Psychology*, 103, 58–72.
- Robinson, E. J., & Apperly, I. A. (1998). Adolescents' and adults' views about the evidential basis for beliefs: Relativism and determinism re-examined. *Developmental Science*, 1, 279–90.

- Samson, D., Apperly, I. A., Braithwaite, J., Andrews, B., & Bodley S. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance* 36(5), 1255–66.
- Samson, D., Apperly, I. A., Kathirgamanathan, U. & Humphreys, G. W. (2005). Seeing it my way: A case of selective deficit in inhibiting self-perspective. *Brain*, 128, 1102–11.
- Sanford, A. J. & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes*, 26, 159–90.
- Saxe, R. (2006) Uniquely human social cognition. *Current Opinion in Neurobiology* 16, 235–9.
- Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, 55, 87–124.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind.” *Neuroimage*, 19, 1835–42.
- Saxe, R., & Powell, L. (2006). It's the thought that counts: Specific brain regions for one component of Theory of Mind. *Psychological Science*, 17, 692–9.
- Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of other's mental states. *Journal of Experimental Psychology: General*, 141, 433–8.
- Sperber, D., & Wilson, D. (2002). Pragmatics, Modularity & Mindreading. *Mind and Language*, 17, 3–23.
- Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, 10, 640–56.
- Surtees, A. & Apperly, I. A. (2012). Egocentrism and automatic perspective-taking in children and adults. *Child Development*, 83(2), 452–60.
- Surtees, A., Butterfill, S., & Apperly, I. A. (2012). Cognitive features of Level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology*, 30(1), 75–86
- van der Meer, L., Groenewold, N. A., Nolen, W. A., Pijnenborg, M., & Aleman, A. (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying Theory of Mind. *Neuroimage*, 56(4), 2364–74.
- Varley, R. & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and ‘theory of mind’ in an agrammatic aphasic patient. *Current Biology*, 10, 723–6.
- Varley, R., Siegal, M., & Want, S. C. (2001). Severe impairment in grammar does not preclude theory of mind. *Neurocase*, 7, 489–93.
- Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N. J., Fink, G. E., & Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage*, 14, 170–81.
- Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: the truth about false-belief. *Child Development*, 72(3), 655–84.
- Wellman, H. M. & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75, 523–41.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–28.
- Zwaan, R. A. & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123, 162–85.
- Zwicker, J. (2009). Agency attribution and visuo-spatial perspective taking. *Psychonomic Bulletin & Review*, 16(6), 1089–93.

Chapter 6

Mind attribution is for morality

Liane Young and Adam Waytz

Morality—judging others’ behavior to be right or wrong, as well as behaving in a right or wrong manner towards others—is an essential component of social life. Morality depends critically on our ability to attribute minds to entities that engage in moral actions (towards ourselves and others) and the entities that experience these actions (our own actions and others’).

The cognitive capacities for attributing minds to others and considering the specific contents of those minds (i.e. mental state reasoning or theory of mind) allow us to understand and interact with individuals and even entire groups of individuals. More specifically, mental state reasoning represents a critical cognitive input for behavior explanation, action prediction, and moral evaluation. We deploy our mental state reasoning abilities in order to explain people’s past actions (e.g. Lisa looked for her shoes in the garage because she *forgot* her mother had moved them to the closet); to predict people’s future behavior (e.g. Mike will tell Barbara his favorite dog joke *not knowing* that Barbara’s dog has just been hit by a car); and to make moral judgments (e.g. Grace must be a bad person for putting what she *thinks* is poison into someone else’s coffee). Our capacity to consider other people’s mental states, including their thoughts, their true or false beliefs, and their helpful or harmful intentions, helps us to navigate our social environment. Indeed, as much research has shown, mental state reasoning functions flexibly across domains, one of which is morality, the focus of this chapter.

The novel claim we make in this chapter is that the *primary* service of mental state reasoning may be for moral cognition and behavior, broadly construed. In particular, the cognitive capacities for mental state reasoning become less relevant when morality is not at stake. We are motivated to understand the actions of relevant moral agents, to predict people’s actions when those actions affect us, directly or indirectly, and to evaluate moral agents as current or future allies or enemies. Computations like these crucially elicit mental state reasoning.

In this chapter, we will therefore review the literature on mental state reasoning for moral cognition—both for judging other moral actors, from the position of “judge” on high, and also for figuring out, as “actors” on the ground, so to speak, who might help us or hurt us, to whom we have moral obligations (for helping or, minimally, not hurting), and whom we ought to trust or avoid (see Figure 6.1).

Morality on high

In this first section, we discuss the critical role of mental states for third-party moral judgments, including how people judge moral agents who harm others. Mental state reasoning is a key cognitive process for evaluating the guilty and innocent intentions of moral agents (Hart, 1968; Kamm, 2001; Mikhail, 2007). Indeed, recent research on the interaction of mental state reasoning and moral cognition has focused on the dominant role of agents’ mental states vs. the outcomes of agents’ actions for our moral judgments (Cushman, 2008; Young, Cushman, Hauser, & Saxe, 2007).

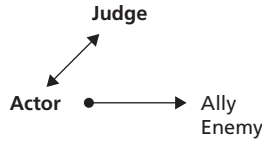


Figure 6.1 Mental state reasoning for moral cognition occurs at multiple levels. Arrows indicate direction of mind attribution. Observers who make third party judgments (“Morality on high”) attribute mind to moral actors. Moral actors who interact with allies and enemies engage in mental state reasoning for affiliation, action understanding and prediction (“Morality on the ground”). Actors may also infer the mind of an evaluative judge (“From the mind on the ground to the mind on high”).

To target the distinct roles of mental states and outcomes, many of these studies present scenarios in which agents produce either a negative outcome (harm to another person) or a neutral outcome (no harm), based on the belief that they would cause the negative outcome (“negative” belief/harmful intention) or the neutral outcome (“neutral” belief/innocent intention). Participants deliver a moral judgment—evaluating the agent’s action as permissible or forbidden, or deciding how much moral blame the agent deserves for his or her behavior.

An example illustrates the possible tension between mental states and outcomes:

Grace and her co-worker are taking a tour of a chemical factory. Grace stops to pour herself and her co-worker some coffee. Nearby is a container of sugar. The container, however, has been mislabeled “toxic”, so Grace thinks that the powder inside is toxic. She spoons some into her co-worker’s coffee and takes none for herself. Her co-worker drinks the coffee, and nothing bad happens.

This scenario pits harmful intentions against neutral outcomes in representing a failed attempt to harm. In an alternative scenario:

A container of poison sits near the coffee. The container, however, has been mislabeled “sugar”, so Grace thinks the powder inside is sugar. She spoons some into her co-worker’s coffee. Her co-worker drinks her coffee and ends up dead.

In this key scenario, an accident occurs—a bad outcome due to a false belief (but not malicious intent). Across studies relying on similar stimuli, participants assigned more moral weight to the agent’s belief and intent, compared to the outcomes (Young et al., 2007). A simple metric of this effect is that participants almost universally judge an attempted harm (e.g. trying but failing to poison someone) as morally worse than an accidental harm (e.g. accidentally poisoning someone).

Other research has investigated not only the simple contrast between intentions and outcomes but also the relative contributions of distinct internal and external factors (e.g. outcome, causation, belief, and desire) for different kinds of moral judgments (e.g. character, permissibility, blame, and punishment) (Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009). Importantly, the agent’s *belief* about whether his or her action would cause harm dominated moral judgments across the board, followed by the agent’s *desire* to cause harm. The relative contribution of beliefs vs. outcomes was greatest for judgments about the moral character of the agent or the moral permissibility of the action. Punishment judgments depended relatively more on outcomes. Nevertheless, these findings indicate the key role of mental state factors for moral judgments.

Notably, mental state factors may underlie moral judgments even in cases where outcomes appear, on the surface, to determine moral judgments. Consider the case of accidents. Many people

assign some blame to agents who cause harmful outcomes, even when they didn't intend to cause the harmful outcomes. (An interesting exception is psychopathy—in the absence of an emotional response to the harmful outcome, psychopaths rely primarily on the stated innocent intent and deliver abnormally lenient judgments of accidents; Young, Koenigs, Kruepke, & Newman, 2012). Recall the scenario in which Grace accidentally poisons her co-worker because she mistakes the poison for sugar. Again, participants mostly excuse Grace on the grounds of her false belief and innocent intention, but they nevertheless assign some moral blame to Grace for the harm done. Behavioral and neural evidence suggests that this moral blame is determined not simply by the harmful outcome of Grace's action; instead, participants' assessment of Grace's mental state drives this judgment (Young, Nichols, & Saxe, 2010b). Participants judge Grace's false belief as more unjustified or unreasonable when it leads to a bad (vs. neutral) outcome, and therefore they judge Grace to be more morally blameworthy. Consistent with this behavioral pattern, activity in brain regions for mental state reasoning, including the right temporo-parietal junction (RTPJ) (Jenkins & Mitchell, 2009; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Saxe & Kanwisher, 2003; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010a), is selectively enhanced when people make moral judgments in response to bad outcomes. In other words, people revise their evaluations of agents' mental states (e.g. whether beliefs were justified or reasonable) in light of the outcome. To summarize, even when we judge accidents harshly, we may do so primarily by considering important mental state factors (e.g. belief justification, negligence, recklessness) and not simply the outcome of the action.

Most of the time, then, internal, unobservable mental states (e.g. beliefs, intentions, desires) carry more moral weight than external outcomes. Extraordinarily, recent research suggests that mental states overwhelm even other external factors, including external, situational constraints (e.g. whether an agent could have done otherwise) (Woolfolk, Doris, & Darley, 2006). In one study, participants read variations of the following story:

Bill discovers that his wife Susan and his best friend Frank have been involved in a love affair. All three are flying home from a group vacation on the same airplane.

In one variation of the story, their plane is hijacked by a gang of ruthless kidnappers who surround the passengers with machine guns, and order Bill to shoot Frank in the head; otherwise, they will shoot Bill, Frank, and the other passengers. Bill recognizes the opportunity to kill his wife's lover and get away with it. He wants to kill Frank and does so. In another variation: "Bill forgives Frank and Susan and is horrified when the situation arises but complies with the kidnappers' demand to kill Frank." When Bill *wanted* to kill Frank, participants actually judged Bill to be more responsible for Frank's death, and the killing to be more morally wrong, even though Bill's desire played no causal role in Frank's death in either case. Mental state factors are clearly at the forefront of our minds when we're making moral judgments.

Blaming immoral agents for their harmful desires and intentions, as in the case of vengeful Bill above, may be easy and automatic for most people (although a key exception, patients with focal lesions to the ventromedial prefrontal cortex (vmPFC), is discussed further below). Forgiving accidents, however, presents a greater challenge. Prior research indicates substantial individual differences among healthy adults in the moral judgments of accidents (Cohen & Rozin, 2001; Sargent, 2004; Young & Saxe, 2009a). In one study, participants who showed greater recruitment of brain regions for mental state reasoning, i.e. the RTPJ, were more likely to forgive accidents, showing greater consideration of the agent's innocent intention (vs. the action's harmful outcome), compared with participants with lower RTPJ responses during moral judgment (Young & Saxe, 2009a).

In development, full forgiveness or exculpation for accidents does not emerge until approximately 7 years of age, surprisingly late in childhood. Interestingly, 5-year-old children appear to be capable of reasoning about false beliefs: in the paradigmatic “false belief task,” children predict that observers will look for a hidden object where they last saw the object and not in its true current location (Flavell, 1999; Wellman, Cross, & Watson, 2001). However, these same children will largely fail to forgive accidents to the same extent as healthy adults: if a false belief leads an agent to unknowingly cause harm to another (e.g. as a result of mistaking poison for sugar), the agent is judged just as bad as though the harm had been caused on purpose (Piaget, 1965/1932). Thus, the ability to integrate mental states (like beliefs and intentions) into moral judgments, vs. the ability to simply encode mental states, may reflect distinct developmental achievements, with distinct functional profiles in the RTPJ (Young & Saxe, 2008). Consistent with this hypothesis, adults diagnosed with Asperger’s Syndrome, who pass standard false belief tasks, also deliver especially harsh moral judgments of accidents (Moran et al., 2011).

Whereas neurotypical adults have particular difficulty exculpating accidents, another population shows a specific deficit in delivering moral judgments of failed attempts to harm, including failed murder attempts—harmful intentions in the absence of harmful outcomes (Young, Bechara, Tranel, Damasio, Hauser, & Damasio, 2010). Patients with focal lesions to the vMPFC judged attempted harms as more morally permissible compared to neurotypical control participants. Strikingly, vMPFC patients even judged attempted harms as more morally permissible than accidents—a reversal of the normal pattern of moral judgments (Cushman, 2008). Consistent with these behavioral data, a recent fMRI study indicates a positive correlation between vMPFC activity and moral judgments of failed attempts to harm; neurotypical participants with high vMPFC responses judged failed attempts more harshly than individuals with low vMPFC responses (Young & Saxe, 2009a). Together, these results suggest that vMPFC patients may be unable to trigger an appropriate emotional response to abstract mental state information, i.e. harmful intentions. The vMPFC may not play a role in encoding mental states *per se*; rather, the vMPFC supports emotional responses to mental state content. This account is consistent with prior work revealing a role for the vMPFC in generating emotional responses to any abstract information (Bechara & Damasio, 2005). Thus, vMPFC patients deliver moral judgments based primarily on the neutral (permissible) outcome, reflecting a “no harm, no foul” mentality.

What, then, are the neural mechanisms that directly support the encoding and integration of mental states in moral judgments? Recent evidence suggests that specific brain regions support multiple distinct cognitive components of mental state reasoning for moral judgment—the initial encoding of the agent’s mental state (Young & Saxe, 2008), the use and integration of mental states (e.g. with outcomes) for moral judgment (Young et al., 2007), spontaneous mental state inference when mental states are not explicitly provided in the scenario (Young & Saxe, 2009b), and even post-hoc reasoning about beliefs and intentions to rationalize or justify moral judgments (Kliemann, Young, Scholz, & Saxe, 2008; Young, Nichols, & Saxe, 2010c; Young, Scholz, & Saxe, 2011).

Building on prior research on the neural substrates for mental state reasoning in the service of action prediction and explanation (Perner et al., 2006; Saxe & Kanwisher, 2003), recent research suggests that a key brain region for moral judgment is the RTPJ. In one study, mentioned above, individual differences in moral judgments were significantly correlated with individual differences in the RTPJ response (Young & Saxe, 2009a). Participants with a high RTPJ response during moral judgment, and a putatively more robust mental state representation (e.g. of the false belief and innocent intention), assigned less blame to agents causing accidental harm. Participants with a low RTPJ response (and weaker mental state representation) assigned more blame, similar to young

children and individuals with Asperger's Syndrome (Moran et al., 2011). One source of developmental change in moral judgments (from a reliance on outcomes to a reliance on mental states) may therefore be the maturation of specific brain regions for representing mental states such as beliefs—consistent with recent research suggesting the RTPJ may be late maturing (Gweon, Dodell-Feder, Bedny, & Saxe, 2012; Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009).

Finally, disrupting RTPJ activity also disrupts the use of mental state information for moral judgment. A recent study probing moral judgments used transcranial magnetic stimulation (TMS) to produce a temporary “virtual lesion” in the RTPJ (Young et al., 2010b). After using fMRI to functionally localize the RTPJ in each participant, offline and online TMS were used to modulate neural activity in two experiments. In both experiments, TMS to the RTPJ vs. the control region reduced participants' reliance on mental states in their moral judgments, and consequently increased the role of outcomes. For example, disrupting RTPJ activity led to more lenient judgments of failed attempts to harm; participants based their moral judgments more on the neutral outcome (vs. the harmful intent). Thus, compromised mental state reasoning in the case of neurodevelopmental disorders (e.g. high functioning autism) or via TMS leads to abnormal moral cognition.

The findings reviewed in this section provide behavioral and neural evidence for mental state reasoning as a key cognitive process for moral judgment. In sum, evaluating moral agents and their actions requires observers to represent and assess the underlying mental states.

Morality on the ground

In this second section, we argue that the key relationship between mind attribution and morality extends beyond the domain of judgment. As social animals, we are not merely passive observers or judges of other people's moral and immoral actions; instead, we are active participants in the social world. We engage in good and bad behaviors toward others, and we must decide how to act toward whom and, in turn, determine who is capable of helping or hurting us. In other words, as moral actors, we must determine who is friend and who is foe. Indeed, the motivation for affiliation with others (e.g. to infer potential allies) and the motivation for action prediction (e.g. to infer potential enemies) are major determinants of mind attribution (Epley, Waytz, & Cacioppo, 2007; Waytz, Gray, Epley, & Wegner, 2010; Waytz, Morewedge, Epley, Monteleone, Gao, & Cacioppo, 2010). It is the *moral* salience of these social contexts that requires and engages mind attribution both for understanding others and for anticipating their actions.

Whether reasoning about allies or enemies, people must engage in mind attribution. Determining who's with us and who's against us (and, at a more basic level, who counts as “us” vs. “them”) through intergroup categorization, is typically an automatic and spontaneous process (Brewer, 1979). Minimal cues to in- and out-group status lead people to encode alliances and coalitions (Kurzban, Tooby, & Cosmides, 2001; Turner, Brown, & Tajfel, 1979). Furthermore, the same neural architecture responds to in- and out-group members after minimal exposure to these individuals. The amygdala, a region involved in processing motivationally relevant information, is responsive to faces of both in-group members and out-group members depending on the processing goals of the perceiver (Lieberman, Hariri, Jarcho, Eisenberger, & Bookheimer, 2005; Van Bavel, Packer, & Cunningham, 2008). Intergroup categorization thus allows us to determine who in our social environment is capable of helping and harming us, and whom we ourselves might be able to help or harm. Thus, allies and enemies alike require social reasoning but elicit distinct motivational strategies. As we argue below, the motivation for affiliation underlies our reasoning about allies, whereas the motivation for action prediction, for anticipating future actions or even attacks, underlies our reasoning about enemies.

The motivation to affiliate with others, and to do good for others, triggers the desire to know others' minds. Understanding the minds of other people is critical for coordination, cooperation, and communication (Epley, & Waytz, 2010). Indeed, a number of research programs have suggested that the capacity for understanding other minds is precisely the capacity that has allowed humans to operate effectively in large social groups (Baron-Cohen, 1995; Humphrey, 1976; Tomasello, Carpenter, Call, Behne, & Moll, 2005). Furthermore, interpersonal liking is often correlated with mind attribution (Kozak, Marsh, & Wegner, 2006), and people will attribute particular mental states, such as secondary emotions, preferentially to in- vs. out-group members (Harris & Fiske, 2006; Leyens et al., 2000). Thus, the motivation for social connection, especially with those within our own moral circle, is a major determinant of mind attribution.

In particular, motivation for social connection leads people to more accurately infer people's emotions from facial or vocal cues (Pickett, Gardner, & Knowles, 2004). This motivation can also increase people's tendency to perceive mental states in non-human entities, such as supernatural agents, technology, and pets, thereby anthropomorphizing them (Aydin, Fischer, & Frey, 2010; Epley, Akalis, Waytz, & Cacioppo, 2008; Epley, Waytz, Akalis, & Cacioppo, 2008). Furthermore, neuroimaging studies have shown that cooperation and generous behavior toward others elicit activity in brain regions that support social cognition including the medial prefrontal cortex (MPFC) (McCabe, Houser, Ryan, Smith, & Trouard, 2001; Waytz, Zaki, & Mitchell, 2012), demonstrating the deployment of mind attribution for positive moral behavior. These findings show that when people seek positive social interactions with other moral agents, they engage in mental state reasoning and may even become hyperattentive to specific features (e.g. emotions) of their social partners' mental states.

Likewise, the motivation to harm others, including our enemies, and to defend against others' harmful actions, also requires a robust understanding of other minds, especially for predicting future actions or attacks. Thus, negative moral interactions are also accompanied by the desire to know others' mental states. As we describe below, the motivation to understand and predict others' actions is therefore another major determinant of mind attribution (Dennett, 1987; Epley et al., 2007).

A number of studies have demonstrated that motivation to attain mastery over others leads to mind attribution. In one instance, this effect obtains for non-human agents; entities that operate unpredictably and that require explanation elicit more attribution of human-like mental states (i.e. anthropomorphism) (Waytz, Morewedge et al., 2010; Morewedge, 2009). When people are motivated to gain control or to explain events in the environment, they will often do so by looking to anthropomorphic Gods or other mentalistic agents (Gray & Wegner, 2010a; Kay, Gaucher, Napier, Callan, & Laurin, 2008; Kay, Moscovitch, & Laurin, 2010; Kelemen & Rosset, 2009). Together, these studies support the idea that the motivation to explain, predict, and understand—the motivation to attain mastery over others—increases mental state reasoning.

Functional neuroimaging evidence suggests that when people are placed in competitive situations with others, in which they must predict and understand others' behavior, brain regions for mental state reasoning including the MPFC (Decety, Jackson, Sommerville, Chaminade, & Meltzoff, 2004) and TPJ (Halko, Hlushchuk, Hari, & Schurmann, 2009) are robustly recruited. One study using positron emission topography (PET) demonstrated that during a competitive game, the MPFC was preferentially engaged when participants believed they were playing an entity capable of strategic moral or immoral behavior (a human being) vs. an entity incapable of such behavior (Gallagher, Jack, Roepstorff, & Frith, 2002). Together, these studies suggest that mind attribution supports not only good moral behavior, such as cooperation with allies, but also strategic interaction with unpredictable others, including enemies.

To demonstrate the relationship between mind attribution and distinct moral motivations towards enemies and allies, we conducted a series of studies targeting both the motivation for social connection and the motivation for action prediction in a single paradigm (Waytz & Young, 2012). In a first study, American participants answered questions about the United States Army and the Taliban, obvious ally and enemy groups, respectively. Participants rated how much they desired social connection with each group and how much they were motivated to predict the actions of each group. Motivation for social connection predicted attribution of mind to the US Army (in-group/ally), whereas motivation for action prediction did not. By contrast, motivation for action prediction predicted attribution of mind to the Taliban (out-group/enemy), whereas motivation for social connection did not. A second study asked American Democrats and Republicans (during the contentious 2010 mid-term elections) to evaluate both the Democratic and Republican party on similar measures, and the same pattern of results emerged. Motivation for social connection uniquely predicted mind attribution toward participants' own political party, whereas motivation for action prediction uniquely predicted mind attribution toward the opposing political party. Taken together, these findings demonstrate that anticipating both positive and negative social interactions (with other moral agents) provokes mind attribution.

Although these dual motivations for effective social interaction engage mind attribution, they may engage different forms of mind attribution. In fact, fMRI research demonstrates that different nodes of the neural network for theory of mind are preferentially engaged by cooperation vs. competition. In one study, in which participants were instructed to play a strategic game, the posterior cingulate was more involved in cooperation, whereas the MPFC was more involved in competition (Decety et al., 2004). Another study demonstrated that reasoning about others' cooperative mental states vs. deceptive mental states recruited distinct brain regions for theory of mind. Whereas both cooperation and deception elicited activation in the TPJ and precuneus, deception selectively increased activation in the MPFC (Lissek et al., 2008). Based on this pattern, the authors suggest that different systems are involved in processing mental states that match an observer's expectations (in this case, cooperative intentions) vs. mental states intended to undermine the observer's expectations. More broadly, these neural findings suggest distinct cognitive processes for mental state reasoning in cooperative vs. competitive contexts.

One hypothesis regarding the differential types of mind attribution for cooperation vs. competition suggests two distinct dimensions of mind. People think about mind in terms of *agency* (i.e. the capacity to plan, to think, and to intend), as well as *experience* (i.e. the capacity to feel pain and pleasure) (Gray, Gray, & Wegner, 2007). The attribution of experience grants a person status as a moral *patient* (i.e. someone who is capable of *experiencing* the moral acts of others), whereas the attribution of agency grants a person status as a moral *agent* (i.e. someone who is capable of *doing* moral acts to others) (Gray & Wegner, 2009; Gray & Wegner, 2010b). Therefore, people express more moral concern toward moral patients, whereas they view moral agents as morally responsible and, therefore, blameworthy or praiseworthy for their actions (Gray et al., 2007).

The tendency to associate experience and agency with distinct moral characters suggests the motivation for social connection, and the motivation for action prediction might differentially trigger attributions of experience and agency, respectively. The desire to give moral care to and receive moral care from another person through prosocial behavior, including cooperation. Therefore, the motivation for social connection should preferentially increase the attribution of *experience* to others. By contrast, the motivation for action prediction entails identifying entities that are capable of planning and acting intentionally and, furthermore, determining the content of those plans and intentions. This motivation should be uniquely linked to the preferential attribution of *agency* to others (Kozak & Czipri, 2011). Two studies demonstrate that when people are tasked with

predicting the actions of a group vs. tasked with affiliating with that group, they prioritize information about the group's capacity for agency vs. experience (Waytz & Young, 2012). Additional behavioral and neural research should uncover whether differential motivations for positive and negative moral interactions map onto the attributions of distinct dimensions of mind.

Most important, considerable research suggests that moral action, and the motivation to engage in moral action—whether positive or negative—depends crucially on mind attribution. People consider the minds of other moral actors not only when judging third-party behavior, but also when attempting themselves to engage with others, either allies or enemies. Behaving well and behaving badly may reside on opposite ends of the moral spectrum, but both depend crucially on mental state reasoning—reasoning about the mind of friends and foes.

From the mind on the ground to the mind on high

In this chapter, we have described how mind attribution is critical for judging moral actions as well as for engaging in good and bad actions towards others. Yet another link between mind attribution and morality, to be explored in future research, is the moral actor's consideration of an evaluative mind or an ultimate judge (see Figure 6.1). A number of studies suggest that when people decide whether to engage in righteous or reproachable actions, they consider whether others are watching, a tendency commonly known as impression management (Leary & Kowalski, 1995). For instance, in monetary exchange games that allow people to behave selfishly or generously, people behave more cooperatively when merely primed with reminders of a judgmental God (Shariff & Norenzayan, 2007) or cues that others are watching (Haley & Fessler, 2005). Perceiving the presence of a mindful, non-human agent also increases honesty and hesitance to cheat in a game (Bering, McLeod, & Shackelford, 2005; Waytz, Cacioppo, & Epley, 2010).

Future research should investigate whether personal decisions about acting morally or immorally, in fact, engage the tendency to search for or perceive a mind on high—either the mind of peer observers or an ultimate moral judge. For now, though, it is clear that mind attribution plays a primary role in both moral judgment and social interactions between moral actors.

References

- Aydin, N., Fischer, P., & Frey, D. (2010). Turning to God in the face of ostracism: Effects of social exclusion on religiousness. *Personality and Social Psychology Bulletin*, 36, 742–53.
- Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge: MIT Press.
- Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, 52, 336–72.
- Bering, J. M., McLeod, K. A. & Shackelford, T. K. (2005). Reasoning about dead agents reveals possible adaptive trends. *Human Nature*, 16, 60–81.
- Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, 86, 307–24.
- Cohen, A. B., & Rozin, P. (2001). Religion and the morality of mentality. *Journal of Personality and Social Psychology*, 81(4), 697–710.
- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition*, 108(2), 353–80.
- Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a “trembling hand” game. *PLoS One*, 4(8), e6699.
- Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & Meltzoff, A. N. (2004). The neural bases of cooperation and competition: an fMRI investigation. *Neuroimage*, 23(2), 744–51.
- Dennett, D. C. (1987). *The Intentional Stance*. Cambridge: MIT Press.

- Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological Science*, 19, 114–20.
- Epley, N., & Waytz, A. (2010). Mind perception. In D. T. G. S. T. Fiske, & G. Lindzey (Eds), *The Handbook of Social Psychology*, 5th edn (pp. 498–541). New York: Wiley.
- Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, 26(2), 143–55.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–86.
- Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, 50, 21–45.
- Gallagher, H. L., Jack, A. I., Roepstorff, A., Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage*, 16, 814–21.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619–19.
- Gray, K., & Wegner, D. A. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, 96(3), 505–20.
- Gray, K., & Wegner, D. M. (2010a). Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review*, 14(1), 7–16.
- Gray, K., & Wegner, D. M. (2010b). Torture and judgments of guilt. *Journal of Experimental Social Psychology*, 46(1), 233–5.
- Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child Development*, 83(6): 1853–68.
- Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26, 245–56.
- Halko, M.-L., Hlushchuk, Y., Hari, R., & Schurmann, M. (2009). Competing with peers: Mentalizing-related brain activity reflects what is at stake. *Neuroimage*, 46(2), 542–8.
- Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low—Neuroimaging responses to extreme out-groups. *Psychological Science*, 17, 847–53.
- Hart, H. L. A. (1968). *Punishment and Responsibility*. Oxford: Oxford University Press.
- Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds), *Growing Points in Ethology*. Oxford: Cambridge University Press.
- Jenkins, A. C., & Mitchell, J. P. (2009). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cereb Cortex*, 20(2), 404–10.
- Kamm, F. M. (2001). *Morality, Mortality: Rights, Duties, and Status*. New York: Oxford University Press.
- Kay, A. C., Gaucher, D., Napier, J. L., Callan, M. J., & Laurin, K. (2008). God and the Government: Testing a compensatory control mechanism for the support of external systems. *Journal of Personality and Social Psychology*, 95, 18–35.
- Kay, A. C., Moscovitch, D. M., & Laurin, K. (2010). Randomness, attributions of arousal, and belief in god. *Psychological Science*, 21, 216–18.
- Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, 111, 138–43.
- Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, 46(12), 2949–57.
- Kozak, M. J., & Czipri, A. (2011). *Behind Enemy Minds: Mind Attribution and Perceived Threat*. New York: Pace University. (Manuscript in preparation).
- Kozak, M. J., Marsh, A. A., & Wegner, D. M. (2006). What do I think you're doing? Action identification and mind attribution. *Journal of Personality and Social Psychology*, 90, 543–55.

- Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(26), 15387–92.
- Leary, M. R., & Kowalski, R. M. (1995). *Social Anxiety*. London: Guildford Press.
- Leyens, J., Paladino, P. M., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A., & Gaunt, R. (2000). The emotional side of prejudice: The attribution of secondary emotions to in-groups and out-groups. *Personality and Social Psychology Review*, *4*, 186–97.
- Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I., & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience*, *8*(6), 720–2.
- Lissek, S., Peters, S., Fuchs, N., Witthaus, H., Nicolas, V., Tegenthoff, M., et al. (2008). Cooperation and deception recruit different subsets of the theory of mind network. *PLoS ONE*, *3*, e2023.
- McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11832–5.
- Mikhail, J. M. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–52.
- Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O’Young, D., Mavros, P. L., et al. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Science USA* *108*: 2688–2692.
- Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General*, *138*, 535–45.
- Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: the roles of left and right temporo-parietal junction. *Soc Neurosci*, *1*(3–4), 245–58.
- Piaget, J. (1965/1932). *The Moral Judgment of the Child*. New York: Free Press.
- Pickett, C. L., Gardner, W. L., & Knowles, M. (2004). Getting a cue: The need to belong and enhanced sensitivity to social cues. *Personality and Social Psychology Bulletin*, *30*, 1095–107.
- Sargent, M. J. (2004). Less thought, more punishment: need for cognition predicts support for punitive responses to crime. *Pers Soc Psychol Bull*, *30*(11), 1485–93.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind”. *Neuroimage*, *19*(4), 1835–42.
- Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Dev*, *80*(4), 1197–209.
- Shariff, A. F., & Norenzayan, A. (2007). God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game. *Psychological Science*, *18*, 803–9.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–91.
- Turner, J. C., Brown, R. J., & Tajfel, H. (1979). Social comparison and group interest in in-group favoritism. *European Journal of Social Psychology*, *9*(2), 187–204.
- Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias: a functional magnetic resonance imaging investigation. *Psychological Science*, *19*(11), 1131–9.
- Waytz, A., Cacioppo, J. T., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*, 219–32.
- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in cognitive sciences*, *14*(8), 383–8.
- Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410.

- Waytz, A., & Young, L. (submitted). *Attributing Mind Across Enemy Lines: When we Treat Outgroups as Mental Agents*.
- Waytz, A., Zaki, J., Mitchell, J. P. (2012). Response of the dorsal medial prefrontal cortex predicts altruistic behavior. *Journal of Neuroscience*, *32*, 7646–50.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Children Development*, *72*(3), 655–84.
- Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283–301.
- Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., & Damasio, A. (2010a). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron*, *65*, 845–51.
- Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010b). Disruption of the right temporo-parietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgment. *Proceedings of the National Academy of Science*, *107*, 6753–8.
- Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–40.
- Young, L., Nichols, S., & Saxe, R. (2010b). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, *1*, 333–49.
- Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*, 1912–20.
- Young, L., & Saxe, R. (2009a). Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–72.
- Young, L., & Saxe, R. (2009b). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*, 1396–405.
- Young, L., Scholz, J., & Saxe, R. (2011). Neural evidence for “intuitive prosecution”: The use of mental state information for negative moral verdicts. *Social Neuroscience*, *6*, 302–15.
- Young, L., Koenigs, M., Kruepke, M., & Newman, J. (2012). Psychopathy increases perceived moral permissibility of accidents. *Journal of Abnormal Psychology*, *121*, 659–667.

Issues in the measurement of judgmental accuracy

David A. Kenny

Introduction

One of the oldest topics of research in the social sciences is the measurement of individual differences in how well it is that people know others. Judgmental accuracy (JA) refers to the ability of people to understand one another. In the prototypical study, a judge is given information about a target (e.g. a photograph or videotape) and asked something about the target. It is then determined if that answer is correct or not. Among the domains studied are the target's personality, opinions, attitudes, moods, emotions, and thoughts. Various other terms for JA have been used—empathy, empathic accuracy, decoding skill, social skills, interpersonal competence, lie and deception detection, interpersonal sensitivity, understanding, and social intelligence to name just some. Simply put JA concerns the ability of a person to understand others. JA is thought to be a one of the four branches of emotional intelligence (Mayer & Salovey, 1997), the perception, appraisal and expression of emotion branch.

JA involves a judge and target. I use in this chapter the term “item” to refer a single rating made by a judge about a given target¹ on a single dimension. The major focus here is on the case in which the judge and target are strangers to each other. With such a constraint, judges have the same information available to them about a given target. When judge and target have a relational history, all sorts of prior information are available. In particular, the judge may have an expectation for the target's typical response.

From the very beginning of research on JA, a key issue is the extent to which there are individual differences in judgmental accuracy. To what extent are some people better at understanding others and others worse? A related question has examined whether particular types of people, e.g. women or leaders, are better judges than others or whether other types of people, e.g. those diagnosed as autistic, are worse judges than others.

This chapter focuses on the measurement of individual differences in JA; that is, the degree to which some persons are consistently better at this task than are others. As will be seen, the measurement of JA is quite difficult and the one goal of this chapter is to understand why it is so difficult. A second goal is to offer suggestions to improve that effort.

Most of the early work on JA focused on self-report inventories, but the current consensus is that judges have little or no insight about their level of JA. For instance, Ickes (1993) stated:

In general, the evidence from the studies my colleagues and I have conducted suggests that people lack metaknowledge regarding their own empathic accuracy. (p. 603)

¹ For some measures of JA, the same target is used more than once, in others only a single target is used, and in others the target is a not one person but more than one person.

Hall, Andrzejewski, & Yopchick (2009) find a weak positive correlation of about Table 7.1 between self-reported ability and actual ability. An alternative, but relatively unexplored, strategy is to use peer ratings of JA (though see a promising exception in Elfenbein, Barsade, & Eisenkraft, 2011).² Both Rosenthal, Hall, DiMatteo, Rogers, & Archer (1979) and Costanzo & Archer (1989) have used peer ratings to validate JA measures. One advantage of peer ratings, over self-ratings, is that one can aggregate ratings over many peers to increase both reliability and validity. However, most of current work in the assessment of JA uses instruments.

Instruments to measure JA can be divided into two basic types: standardized instruments (e.g. the PONS (Rosenthal et al., 1979) or the CARAT (Buck, 1976)) and standardized formats (e.g. Ickes empathic accuracy paradigm (1993) or Buck's slide viewing technique (Sabatelli, Buck, & Kenny, 1986)). A standardized instrument is based on the logic of a personality or intelligence test. The same items are used for all judges taking the test. In a standardized format, the basic task is the same, but the targets are specially chosen for the judges.

The focus in this chapter is on standardized instruments to measure JA. Several different investigators had developed a standardized instrument to measure judgmental accuracy. For this chapter,³ I have selected a heterogeneous, but clearly non-random collection of eight of these measures:

- ◆ *Communication of Affect Receiving Ability Test* (CARAT; Buck, 1976): 30 spontaneous facial video clips of adult targets watching four categories of emotionally evocative slides—scenic, pleasant people, unpleasant, and unusual.
- ◆ *Diagnostic Analysis of Non-verbal Accuracy Scale* (DANVA or DANVA-2-AF; Nowicki & Duke, 1994): 24 posed photographs of adult facial expressions of six targets and four emotions—happiness, sadness, anger, and fear.
- ◆ *Eyes* (Reading the mind in the eyes; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001): the revised test with 36 items, each a picture of the eyes, with four response alternative (e.g. playful, comforting, irritated, or bored).
- ◆ *Interpersonal Perception Task 30* (IPT30; Costanzo & Archer, 1989): 30 audiovisual excerpts of individual targets talking or small groups of targets interacting. Content covers intimacy, kinship, competition, status, and deception.
- ◆ *Interpersonal Perception Task 15* (IPT15; Archer & Costanzo, 1993): the 15 “best” audiovisual excerpts selected from the IPT30.
- ◆ *Profile of Non-verbal Sensitivity* (PONS; Rosenthal et al., 1979): a 220-item test containing 2-second clips of all combinations of face, body, electronically filtered speech, and random-spliced speech (total of 11 channels) of an adult female target portraying 20 different affective situations.
- ◆ *Sternberg & Smith 1* (S&S1; Sternberg & Smith, 1985): a 70-item test that presents a picture of two people and the judge is asked if the two are in a relationship or not.
- ◆ *Sternberg & Smith 2* (S&S2; Sternberg & Smith, 1985): a 70-item test that presents a picture of two people and the judge is asked which of the two people is the subordinate and which is the supervisor.

² It is interesting to note that emotional intelligence branch that shows the weakest level of peer agreement (about 8% of the variance in two studies) is the perception, appraisal and expression of emotion branch, the one emotional intelligence branch that is closest to JA.

³ The survey of instruments is skewed toward older and more “classical” measures. Although more modern measures have somewhat higher reliability than do the measures, they are, nonetheless, still much lower than what most researchers expect.

Table 7.1 Number of items (*k*), the proportion of items correct of (*P*), number of alternative answers (# *A/t*), Cronbach alpha reliability (α), average inter-item correlation ($r_{1,1}$), and reliability with 24 items ($r_{24,24}$) for eight selected measures of judgmental accuracy^a

Test	<i>k</i>	<i>P</i>	# <i>A/t</i>	α	$r_{1,1}$	$r_{24,24}$
CARAT	30	.600	4	.460	.028	.405
DANVA	32	.889	4	.880	.186	.846
Eyes	36	.728	4	.488	.026	.389
IPT-30	30	.557	2 & 3	.290	.013	.246
IPT-15	15	.631	2 & 3	.240	.021	.336
PONS	220	.773	2	.860	.027	.401
S&S I	41	.600	2	.620	.038	.489
S&S II	36	.740	2	.590	.038	.490

^aSee Appendix A for more information on measures and the sources used.

Appendix 7A contains more details about each of the measures and where the numeric values that are reported in this chapter come from. Should the reader want to take one of these tests, an online version of the Eyes test is available at <http://glennrowe.net/BaronCohen/Faces/EyesTest.aspx>.

Historically, these standardized measures are plagued by issues of very low internal consistency reliability. For instance, after reviewing the literature on the poor reliability of measures of JA, Kenny & Albright (1987) noted (p. 393)

Our position is not that individual differences are nonexistent in interpersonal accuracy. Rather, we believe that the variability of such differences is rather limited.

More recently, Hall, Halberstadt, & O'Brien (1997, p. 302) noted:⁴

Consistent with prior experience ..., the nonverbal decoding tests showed weak internal consistency. Cronbach's alpha coefficients for the male and female subtests of the IPT, CARAT, and PONS ranged from .10 to .35, with a median of .21.

Alphas of .21, unheard of in most areas of social science, are quite commonplace in the study of JA.

Table 7.1 presents the internal consistency estimates of reliability of measurement for eight different measures of JA. Appendix 7A gives what studies were used to compute the reliability estimate. It should be noted that sometimes (e.g. CARAT, IPT15, and IPT30), the original reliabilities were much too optimistic. No doubt in the initial selection of items, items were chosen because of higher inter-item correlations. If there was such capitalization on chance, the expectation would be for lower correlations in subsequent samples. For these measures, a more recent, and likely more representative value is used. Table 7.1 also presents *P*, which is defined here as the mean proportion of the items that were answered correctly. To help interpret this value, the number of alternatives for the test is given.

From the reliabilities, using the Spearman–Brown prophecy correlation, the average inter-item correlation, $r_{1,1}$, is computed, which is the correlation of one item with one other item and can be viewed as the reliability of a single item. Because the $r_{1,1}$ correlation does not depend on the number

⁴ The PONS measured used is the short form.

of items, the $r_{1,1}$ correlations are more comparable across measures with a different number of items. Also given in Table 7.1 is the $r_{24,24}$ which is a forecast of the test's reliability if it had 24 items. This column gives a forecast of each test's reliability with same number of items.

In terms of the $r_{1,1}$, it is seen that the median value is about .03 which leads to a reliability for a scale with 24 items of .390. How does an inter-item correlation of .03 compare with other psychometric tests? The answer is not very well. In terms of cognitive ability tests, for the Peabody Picture Vocabulary Test, the value of $r_{1,1}$ is .085 and for the Raven's Matrices test it is .109. So it is seen that these values are about three times higher than for most measures of JA. Self-report measures are even higher, with the Rosenberg self-esteem scale at .340, for Beck Depression Inventory at .300, and for the Bem Masculinity and Femininity scale at .190. Clearly, the measures of $r_{1,1}$ for JA are much lower than what are typically found in psychological measurement.

Some caution needs to be taken in interpreting these low reliability measures. Certainly, many of these internal consistency estimates of reliability are very low and well below the generally agreed-upon minimal standard of .70. However, these values do not indicate that measures of JA have *no* reliability. Some measures of JA, but not all, have poor reliability, but still about half of their variance is reliable. An important lesson to be learned is that JA is very difficult to measure, something that should be confronted and not ignored.

Alternative measures of reliability

Given these low internal consistency reliabilities, some have argued that internal consistency formulas are inappropriate in the measurement of JA. The lack of enthusiasm for internal consistency measures of reliability is hardly surprising with one study (Mayer & Geher, 1996) reporting a reliability of only .24 for a 96-item scale!

One alternative proposed is a test-retest correlation. However, as discussed by Gignac (2009) a test-retest correlation can be a very misleading measure of reliability:

Consider the three following variables: height, intelligence, and extraversion. The correlations between these variable measured in adults would all be expected to be less than .20, which would preclude any meaningful aggregation of the scores ... However, these same aggregated scores would be expected to have very high levels of test-retest reliability, because all three of the scores would not be expected to change. (p. 22)

Although it may be comforting to know that the measure is stable, test-retest measures are not very informative measures of reliability.

It has also been argued that because JA is a multidimensional construct (Schlegel, Grandjean, & Scherer, 2011) and, as such, an internal consistency measure is inappropriate. From a multidimensional inventory, a parallel forms measure of reliability would be the appropriate measure of reliability. To compute a parallel forms reliability, in principle a second form of the test would need to be created and correlated with the original test. Practically, a split-half reliability is computed as follows. As a hypothetical example, presume that a judge's skill at reading a target's non-verbal skill is due to three different skills—skill at reading the face, body, and voice. These different skills might well be weakly and perhaps even negatively correlated. However, within-test items that only contain the face there should show good internal consistency. To compute a split-half reliability, the items need to be divided into three sets containing items from each of the dimensions, face, body, and voice. Then each of these sets is split in half, form one and form two, and then the sum of the items from form one is correlated with the sum from form two. This correlation is then adjusted by Spearman–Brown prophecy formula for doubling a test to obtain a measure of reliability. So if the scale had 10 face items, eight body items, and 12 voice items, each form would have five face items, four body items, and six voice items.

Parallel forms reliability lead to larger reliability values than internal consistency, but often not a lot larger. For instance for a 40-item test with 10 dimensions, each with four items, a .05 inter-item correlation within dimension and a .02 between items for different dimensions (the average latent correlation between dimensions being .4), the split-half reliability would only be .009 greater than the internal consistency reliability (.659 vs. .650). Not only are the benefits not as great as might be thought, there are significant costs involved in adopting this strategy. First, some sort of theory dimensionality must be available to classify the items into different types. That is, one needs an *a priori* theory to divide the items up into different types. Second, causal assumptions become more complicated. If the researcher hopes to show that his or her measure of JA is a cause of a given outcome or is caused by a given antecedent, it is important to show that the effect is for the overall construct and not just the separate components. Returning to the example of non-verbal ability with three dimensions (face, body, and voice), if gender differences were found, it would be important to know if that gender difference occurred in all three dimensions, and if non-verbal ability predicted interpersonal success, it would be important to know if all three dimensions did so.

Even if a split-half reliability is computed to determine reliability, it is still necessary that items with each dimension show internal consistency. Demonstrating replicability of measurement is key to establishing a measure's reliability.

Validity of JA

Reliability, although important, is not the most important attribute of a JA measure. What is of the essence is validity. Reviewed here, very briefly, is evidence of convergent (do two measures of the construct correlate with each other?) and predictive validity (does the measure correlate with other constructs associated with the theoretical construct).

In terms of convergent validity, the different tests of JA correlate very weakly at best. Based on prior analyses of Hall (2001), the correlation between different measures of JA tend to be very low, perhaps averaging about .10. With such small correlations, sometimes even a negative correlation is found. In fact the two Sternberg & Smith (1985) measures correlate $-.09$ with each other. The low convergent validity points to multi-dimensionality in the measurement of JA, an issue discussed above when parallel forms reliability was discussed.

It can also be asked if measures of JA have predictive validity. If measures consistently correlated a given variable in a meaningful way, then there would be evidence that the measure is tapping what it purports to be measuring. In the first ever meta-analyses published in *Psychological Bulletin*, Hall (1978) found evidence for gender differences nonverbal decoding ability, the average correlation being .20 with females scoring higher than men. Hall et al. (2009) found an average correlation of $-.04$ for masculinity and .12 for femininity. In a large meta-analysis of 206 independent studies, Hall et al. (2009) found consistent, but positive personality traits ($r = .08$), negative personality ($r = -.07$), self-rated social competence ($r = .10$), and other-rated social competence ($r = .07$). Although many of the effect sizes are small at best, the pattern of results is consistent. JA measures do correlate meaningfully and consistently with the variables that they should correlate with.

Strategies for improving the reliability of measurement of JA

Presented here are three ways for improving the reliability and in principle the validity of measures of JA. Not discussed, but a very successful strategy is to use that of the PONS—use a lot of items. However, note the DANVA has just 24, only about one-ninth as many as the PONS, but even greater reliability.

“Easy” tests

Previous analyses of JA have used classical test theory (CTT). That model states that the observed score equals a true score plus error. A more appropriate model is item response theory (IRT), which is used to describe the probability that the judge makes the correct response. In the classical IRT model, it is assumed that the response is dichotomous, correct or incorrect, and not continuous as assumed by CTT. Moreover, IRT has an explicit model of guessing.

Appendix 7B outlines the IRT model that was used in the simulation. A measure of JA is assumed to contain 24 equally discriminable and difficult items. Varied is the true variance, or σ , as either .5 or 1.0. The α reliability of 24-item test, assuming no guessing, would be about .75 for $\sigma = 1.0$ and .53 for $\sigma = .5$. Also assumed that if the participant does not know the correct answer, a random guess is made and it is assumed that there are just two alternatives. The one parameter that is varied is across trials is the average difficulty of the items. A very difficult test would have as the average proportion correct or P near .5 (not zero due to guessing) and a very easy test would have a probability near one. The usual view is that reliability should be near its maximum when the average probability is about half way between chance and perfect or .75 when there are two alternatives. As seen in Table 7.1, many JA tests have as the average number correct about .75 (assuming just two alternatives). The following quote from Hall, Andrzejewski, Murphy, Schmid Mast, & Feinstein (2008) reveals the logic of having a test with P equal to .75:

(T)he PONS test was designed to have a proportion accuracy of about .75 (midway between the guessing level of .50 and perfect accuracy of 1.00), in the belief that scores midway between guessing and perfect accuracy would be optimal for revealing individual differences. (p. 1479)

Thus, we might expect that Cronbach's alpha would maximize when the average is near .75. However, as seen in Figure 7.1, alpha reliability does not maximize at .75, but at a value higher than that. Moreover, as explained in Appendix B, IRT predicts that alpha would maximize for values of 0.880 for $\sigma = 1$ which corresponds to P of .853 and equals 0.640 for $\sigma = .5$ which corresponds to a P of .827. Interestingly and importantly, it should be noted that the DANVA, which has the best reliability, has by far the largest probability correct.

Why does this happen? Consider what would happen when the average probability correct or P is .75. Consider two people who both have low ability, Jack scoring 1.5 standard deviations below the mean and Jill scoring 2.0 standard deviations below the mean. Jack has a 53.6% chance of getting

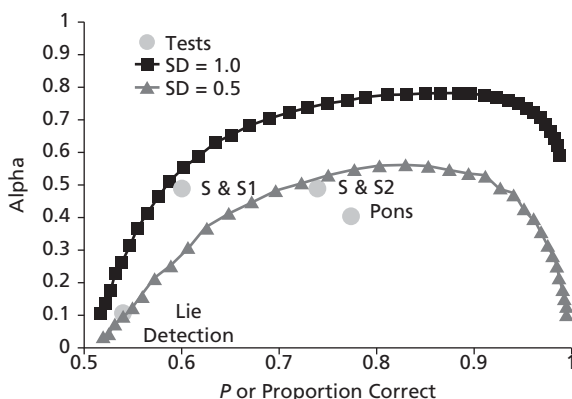


Figure 7.1 Alpha (assuming 24 items) as a function of the proportion correct.

an item correct whereas Jill has only a 51.6% chance of getting an item correct. Jack should do better on a test than Jill, but luck would play a big role here. For Jack to have at least a 75% chance of outscoring Jill, the test would need to be about 567 items! If, however, the test was changed and the probability of someone of average ability getting the item correct was raised to .88, Jill has about a 54.7% of getting an item correct, whereas Jack has a 59.9% chance. Now for Jack to have at least a 75% chance of outscoring Jill, there would need to be about 83 items. Luck would play a much smaller role in determining who is better if the test is “easy.”

The literature does suggest that there is a relationship between how easy a JA test is and how reliable it is. At one end is lie detection where the mean level is about .54 and individual differences are very weak (Bond & DePaulo, 2006). At the other end is the DANVA, which is relatively easy with a mean level of .89 and high levels of reliability. Figure 7.1 also contains the forecasted reliabilities for each of the standardized tests with two response alternatives, as well as lie detection, under the assumption that the test has 24 items. We see that all of the values fall near the $SD = 0.5$ which indicate weak reliability. (The fact that S&S1 is higher is likely due to capitalization on chance due to dropping many of the items.)

It is also noted that the PONS subscale with the lowest alpha reliability (randomized-spliced audio with no video) has the lowest mean score ($P = .627$), whereas the two subscales tied with largest reliability (random-spliced audio with face video and content-filtered audio with figure video) have the two highest mean scores ($P = .884$ and $.853$, respectively; see Tables 3.1 and 3.5 in Rosenthal et al., 1979). This illustrates that test reliability could be enhanced greatly if tests were made easier. Finally, across the 11 subtests of the PONS, the correlation of P with the reliability is .82.

It should be pointed out that if tests are easier, it would probably be easier to increase the number of items as the items would likely be completed more quickly. Thus, it would be possible to use the strategy developed for the PONS of having a large number of items to increase reliability. Additionally, it should be pointed out that a very “easy” test would not be very useful at discriminating ability differences of those who are very skilled at JA.

Using the psychometric correct answer

There has been considerable debate around the issue of what is considered to be the correct answer in the JA. The first and most obvious choice is to use the truth to determine the correct answer. However, some standardized instruments use posed emotions (e.g. PONS) making the meaning of the truth unclear. In this case, it is presumed that the target is posing the proper emotion. Particularly for cases in which there is no clear true answer, researchers have used two different strategies. One idea is to use an expert opinion, usually the researcher. The other idea is to use what is called a *consensus answer*, the answer that a plurality of respondents choose.

One alternative that appears to not be used is what might be called the *psychometric truth*. It is assumed that all the items measure the same construct, if only weakly. So the “truth” is determined by a psychometric standard, e.g. the correlation of the item with the mean of the other items, commonly called the item-total corrected correlation. If the item has a non-trivial, negative item-total corrected correlation, the scoring of the item would be reversed, making the false answer now the true answer. By using such a scaling, one would improve the reliability of measurement.

Such a strategy has likely not been previously adopted or even recommended because it seems nonsensical—why score an item where the wrong answer is treated as the right answer? The reason is that sometimes doing the right thing can lead to a bad result, but that does not mean that doing the right thing is wrong. As an over-simplified example, imagine if judging happiness the only relevant cue is whether the target is smiling or not. If someone is smiling, it is much more likely that

the person is happy. Imagine further that for some targets, they may be smiling, but they are not happy. For these targets, a judge who used the cue of smiling would be wrong, but the judge would be wrong by using the right strategy.

I realize that is a controversial suggestion, but it would be relatively easy to check. What is needed is a test with many participants and some sort of criterion measure to assess validity. First, it would be determined through the item-total corrected correlations which items need to be reversed. Then the sum of those items, unreversed, should correlate negatively with the criterion measure.

Embodied JA

Measuring JA using a standardized instrument such as the PONS or DANVA has some advantages but it has disadvantages. The two major advantages are first that it is relatively easy to administer a standardized test and that because all judges view the same set of items, individual differences represent a pure measure of decoding ability, i.e. receiving ability. The major disadvantage is that the judgment task is decontextualized. The judges are not given any meaningful context about the behaviors that they observe. Moreover, the target is a total stranger to the judge, they have no social interaction history, and they are not interacting with target when the behavior occurs. Everyday person perception is embodied and unfolds over time in social-interaction contexts in which the judge and target have goals and a history. Wilhelm & Perrez (2004) have shown that judges take advantage of this knowledge, and they know how their partner typically responds, they can assume that their partner is responding as they would respond, and they can even predict their partner's emotional response when they are not together.

Certainly embodied person perception is more difficult to study, but those difficulties may well be worth it to reveal individual differences in person perception. One suggestion is to employ a standardized format with different targets. So for instance, using a standardized format, how accurate someone is at reading fellow group members can be measured. A useful tool with standardized formats is the Social Relations Model (Kenny, 1994). Within that model, a set of judges rates multiple targets who may well also be judges. In these analyses, "item" is the target, and the consistency of judges can be measured across targets: If Sam is a good judge of Jack, is he also a good judge of Jill? An estimate of that proportion is .06 (see Appendix 7C), which can be viewed as an inter-item correlation or $r_{1,1}$ where the item is target. Such a value is somewhat higher than the inter-item correlation using standardized instruments, but not a lot higher. The reliability of judgment can be measured, assuming a fixed number of targets. For example, with 10 targets, that reliability is .39. Biesanz (2010) has extended the Social Relations Model to account for multivariate responses for each target.

There are two final points about measuring accuracy in a more naturalistic context I want to make. First, successful person perceivers realize that the target has changed, which requires a research design in which the judge observes the target over time (Neyer, Banse, & Asendorpf, 1999). Secondly, it is important to realize that accuracy in person perception often occurs through the use of heuristics, e.g. assumed similarity. Recent work by West & Kenny (2011) captures how it is that judges can be both biased and accurate.

Conclusion

The topic covered in this chapter is extensive, and so many key topics in the measurement of JA are not considered here. One of the most important is that guessing is not usually random. The classic paper by Wagner (1993) emphasized the importance of base rates in the response measures, something not considered in this chapter.

The chapter has documented the apparent contradiction that standardized measures of JA have low reliability yet have some degree of validity. Certain partial remedies have been suggested, but further work is needed. The hope would be that increased knowledge about JA, as well as the application of psychometrics, would lead to further improvements in the measurement of JA. Being able to understand others is a remarkable skill that humans possess and it is then important that science can measure that skill.

Acknowledgements

I want to thank Judith Hall who earlier provided me with advice on several key issues and made numerous helpful comments on a prior draft. Katja Schlegel also provided me with several valuable comments. Finally, I want to acknowledge that much of the work on this chapter was undertaken when I received an Erskine Fellowship from Christchurch University, Christchurch New Zealand.

References

- Archer, D., & Costanzo, M. (1993). *The Interpersonal Perception Task-15 (IPT-15)*. Berkeley: University of California Extension Media Center.
- 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*, 42, 241–51.
- Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research*, 45, 853–85.
- Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Bulletin*, 10, 214–34.
- Buck, R. (1976). A test of nonverbal receiving ability: Preliminary studies. *Human Communication Research*, 2, 162–71.
- Costanzo, M., & Archer, D. (1989). Interpreting the expressive behavior of others: The Interpersonal Perception Task. *Journal of Nonverbal Behavior*, 13, 225–45.
- deAyala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.
- Elfenbein, H. A., Barsade, S., & Eisenkraft, N. (2011). *Do We Know Emotional Intelligence When We See It? The Properties and Promise of Observer Ratings*. Unpublished paper, Washington University, St Louis.
- Elfenbein, H. A., Foo, M. D., Boldry, J. G., & Tan, H. H. (2006). Dyadic effects in nonverbal communication: A variance partitioning analysis. *Cognition and Emotion*, 20, 149–59.
- Gignac, G. E. (2009). The psychometrics in the measurement of emotional intelligence. In Stough, C., Saklofske, D. H., & Parker, J. D. A. (Eds), *Assessing Emotional Intelligence: Theory, Research, and Applications* (pp. 9–40). New York: Springer.
- Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin*, 85, 845–57.
- Hall, J. A. (2001). The PONS test and the psychometric approach to measuring interpersonal sensitivity. In J. A. Hall and F. J. Bernieri (Eds), *Interpersonal Sensitivity: Theory and Measurement* (pp. 143–60). Mahwah: Erlbaum.
- Hall, J. A., Andrzejewski, S. A., Murphy, N. A., Schmid Mast, M., & Feinstein, B. (2008). Accuracy of judging others' traits and states: Comparing mean levels across tests. *Journal of Research in Personality*, 42, 1476–89.
- Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior*, 33, 149–80.
- Hall, J. A., Halberstadt, A. G., & O'Brien, C. E. (1997). "Subordination" and nonverbal sensitivity: A study and synthesis of findings based on trait measures. *Sex Roles*, 37, 295–317.

- Ickes, W. (1993). Empathic accuracy. *Journal of Personality*, **61**, 587–610.
- Ickes, W., Buysse, A., Pham, H., Rivers, K., Erickson, J. R., Hancock, M., et al. (2000). On the difficulty of distinguishing “good” and “poor” perceivers: A social relations analysis of empathic accuracy data. *Personal Relationships*, **7**, 219–34.
- Kenny, D. A. (1994). *Interpersonal Perception: A Social Relations Analysis*. New York: Guilford.
- Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*, **102**, 390–402.
- Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*, Vol. 18 (pp. 142–82). Orlando: Academic.
- Malone, B. E., & DePaulo, B. M. (2001). Measuring sensitivity to deception. In J. A. Hall & F. J. Bernieri (Eds), *Interpersonal Sensitivity: Theory and Measurement* (pp. 103–24). Mahwah: Erlbaum.
- Mayer, J. D., & Geher, G. (1996). Emotional intelligence and the identification of emotion. *Intelligence*, **22**, 89–113.
- Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. Sluyter (Eds), *Emotional Development and Emotional Intelligence: Implications for Educators* (pp. 3–31). New York: Basic Books.
- Neyer, F. J., Banse, R., & Asendorpf, J. B. (1999). The role of projection and empathic accuracy in dyadic perception between older twins. *Journal of Social and Personal Relationships*, **16**, 419–42.
- Nowicki, S., Jr., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior*, **18**, 9–35.
- Patterson, M. L., Foster, J. L., & Bellmer, C. (2001). Another look at accuracy and confidence in social judgments. *Journal of Nonverbal Behavior*, **25**, 207–19.
- Patterson, M. L., & Stockbridge, E. (1998). Effects of cognitive demand and judgment strategy on person perception accuracy. *Journal of Nonverbal Behavior*, **22**, 253–63.
- Ragsdale, G., & Foley, R. A. (2011). A maternal influence on reading the mind in the eyes mediated by executive function: Differential parental influences on full and half-siblings. *PLoS One*, **6**, e232360.
- Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to Nonverbal Communication: The PONS Test*. Baltimore: Johns Hopkins.
- Sabatelli, R. M., Buck, R., & Kenny, D. A. (1986). A social relations analysis of nonverbal communication accuracy in married couples. *Journal of Personality*, **53**, 513–27.
- Schlegel, K., Grandjean, D., & Scherer, K. R. (2011). Emotion recognition: Unidimensional ability or a set of modality- and emotion-specific skills? *Personality and Individual Differences*, **53**, 16–21.
- Sternberg, R. J., & Smith, C. (1985). Social intelligence and decoding skills in nonverbal communication. *Social Cognition*, **3**, 168–92.
- Thomas, G., & Fletcher, G. J. (2003). Mind-reading accuracy in intimate relationships: Assessing the roles of the relationship, the target, and the judge. *Journal of Personal and Social Psychology*, **85**, 1079–94.
- Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, **17**, 3–28.
- West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review*, **118**, 357–78.
- Wilhelm, P., & Perrez, M. (2004). How is my partner feeling in different daily-life settings? Accuracy of spouse's judgements about their partner's feelings at work and at home. *Social Indicators Research*, **67**, 183–246.
- Woods, E. (1996). Associations of nonverbal decoding ability with indices of person-centered communicative ability. *Communication Reports*, **9**, 13–22.

Appendix 7A: Details about sources used for entries in Table 7.1

- ◆ CARAT: Buck (1976) reported a reliability of .56. Used here is the value reported by Hall (2001, p. 152), which is lower than the Buck's original value. The mean is taken from Buck (1976).
- ◆ DANVA (*Diagnostic Analysis of Non-verbal Accuracy Scale*): this is the Faces-Receptive scale described by Nowicki & Duke (1994). The mean was computed from Table 2 and the reliability from Table 3.
- ◆ *Eyes (reading the mind in the eyes)*: the alpha reliability was computed from Baron-Cohen et al. (2001) using means and standard deviations of the normal sample. Interesting, a very similar alpha reliability of .481 is given in Ragsdale & Foley (2011). The mean was also taken from the normal sample in Baron-Cohen et al. (2001).
- ◆ IPT 15: Archer & Costanzo (1993) report a reliability of .38. Used here is the average from Patterson, Foster, and Bellmer (2001), three values reported by Hall (2001), and Woods (1996). The mean is taken from Archer & Costanzo (1993).
- ◆ IPT 30: Costanzo & Archer (1989) report a reliability of .52. Used here is the average of the values reported by Hall (2001) and Patterson and Stockbridge (1998). The mean is taken from Costanzo & Archer (1989).
- ◆ PONS: The reliability and means are taken from Tables 3.1 and 3.5 of Rosenthal et al. (1979).
- ◆ S&S1: the mean is taken from Table 1 of Sternberg & Smith (1985) and the item reliability from page 180. Note that 29 items were dropped from the scale because of low item total correlations. Because the scale was not cross-validated, likely the reliability is inflated. The reliability of the full 70 item scale is .34.
- ◆ S&S2: the mean is taken from Table 1 of Sternberg & Smith (1985) and the item reliability from page 180. Note that 34 items were dropped from the scale because of low item total correlations. Because the scale was not cross-validated, likely the reliability is inflated. The reliability of the full 70-item scale is .47.

Appendix 7B: Details concerning the IRT model of JA and simulation

In the IRT model, the ability of an individual i to know the answer to a question is denoted here as r_i where the subscript will be dropped for simplification. It is assumed that all items are equally difficult (akin to item means) and equally sensitive, where the sensitivity parameters (akin to factor loadings) are set to one. Although these are very strong assumptions, relaxing them does not substantially change the conclusions that follow. It is assumed that ability or r has a normal distribution with a mean of zero and a variance of σ . In the simulation, σ is set to 0.5 and 1.0. As is the usual convention in IRT, the value of r is multiplied by 1.7 to approximate as logistic distribution.

In this simulation, the item difficulty parameter, assumed to be constant across items for any given trial, is f where $1 - e^f/[1 + e^f]$ (e being the irrational number that approximately equals 2.718) gives the probability that someone of average ability ($r = 0$) knowing the correct answer. The larger the value of f , the more difficult the test. The probability of someone with average ability knowing the correct answer is given by $e^f/(1 + e^f)$. Note that if f is zero, then this probability is .5. Within the IRT model, the probability of a person with ability r knowing the correct answer is given by $e^{r^f}/(1 + e^{r^f})$ (and the subscript i is dropped for r).

In IRT, if someone does not know the correct answer, the person guesses. Two alternatives are assumed and it is assumed that the probability of being correct is assumed to be .5. Guessing can be added to the model by what is called in IRT as the three-parameter model and the probability of being correct equals:

$$\frac{e^{r-f}}{1 + e^{r-f}} + 0.5 \left[1 - \frac{e^{r-f}}{1 + e^{r-f}} \right] \quad (\text{A1})$$

The predicted proportion correct of someone of average ability ($r = 0$) or P is:

$$0.5 \left[\frac{1 + 2e^{-f}}{1 + e^{-f}} \right] \quad (\text{A2})$$

Thus, when f equals 0, P equals .75.

The model in Equation A1 was estimated with 24 items and the standard deviation of individual differences or σ was either 0.5 or 1.0. The value of Cronbach's alpha, assuming no guessing, would be .75 for $\sigma_s = 1.0$ and .53 for $\sigma_s = 0.5$. In the simulation the parameter f , the item difficulty parameter, was allowed to range from 5 to -5 in increments of .25. A total of 100 trials each with 500 cases were run. For each trial we saved P , the average proportion correct and Cronbach's alpha. The results are shown in Figure 7.1 and show that alpha maximizes at a value greater than P equal to .75.

Within IRT, reliability is measured by a parameter called *information*. Following deAyala (2009, p. 144), the information maximizes at the point when f (the item difficulty parameter in IRT) equals:

$$-.693 + \sqrt{0.5 + 6.8\sigma} \quad (\text{A3})$$

This value is called the offset. Note that Equation A3 equals 0.880 for $\sigma = 1$, which corresponds to P of .853 and equals 0.640 for $\sigma = .5$, which corresponds to a P of .827. Note that both of these values are considerably larger than .75. These two values correspond closely to the values obtained in our simulation (see Figure 7.1).

The model could be complicated to allow some items to be more diagnostic than other items (akin to larger factor loadings in CTT) and for item difficulty to vary. However, the major point that is made in this section that the optimal value of P is greater than .75 is not affected by these assumptions.

Appendix 7C: Social relations variance partitioning of JA

With the Social Relations Model (SRM), the set of judges rate the same set of targets, who may very well be the same persons as the judges. For each judge-target pair, an accuracy score is computed. Within the SRM, the variance in accuracy is partitioned into the following sources:

Judge: Are some judges better than others at the task?

Target: Are some targets easier to judge than others?

Relationship: Are people better at judging some persons more than others, controlling for judge and target effects?

Error: Unpredicted variance.

Table 7.A3 Social relations variance partitioning and reliability values from three studies

Study	Judge	Target	Relationship/ error	$r_{1,1}$	Reliability ^a
Kenny & La Voie (1984)	.03	.42	.55	.05	.35
Ickes et al. (2000)	.00	.33	.67	.00	.00
Elfenbein et al. (2006)	.09	.28	.63	.13	.59
Average	.04	.34	.62	.06	.39

^aAssuming 10 targets.

Without replications, relationship and error variance are confounded. Most studies have not attempted to measure group variance.

In Table 7.A3 are the proportions of judge, target, and relationship/error variance. The following studies are included;

Kenny & La Voie (1984): The average across three studies of emotion and deception detection are presented.

Ickes, Buysse, Pham, Rivers, Erickson, Hancock, et al. (2000): The average across the two studies of empathic accuracy that do not use standardized stimuli.

Elfenbein, Foo, Boldry, & Tan, (2006): One study involving the perception of emotion.

Also reported in the table is (judge variance)/(judge variance + relationship/error variance) which is a close analogue to the inter-item (target) correlation or $r_{1,1}$ and the reliability of judgment assuming ten targets.

Not directly relevant to this paper is the interesting result that target is the dominant source of variance: Some targets are easy to judge and others are most difficult. The statement by Malone & DePaulo (2001) is quite relevant here:

It is possible that most of the variance ... is due to differences in the judgeability of targets as opposed to the sensitivity of the perceivers. (p. 113)

Not included in the table is the study by Thomas & Fletcher (2003) which has quite different results, yielding $r_{1,1}$ of .62 and an internal consistency estimate with 10 targets of .94. The large value found in this study, but perhaps it is due to the fact that judges were placed in a highly emotional situation.

Section 2

Neural systems and mechanisms

This page intentionally left blank

Brain electrophysiological studies of theory of mind

Mark A. Sabbagh

For nearly 100 years, brain encephalographic recordings (EEG) have provided a way of monitoring brain activity to gain a window into both the overall condition of the brain, and the brain's contribution to cognitive activity (see Millett, 2001 for a review). It is currently widely used as a relatively low cost (per subject), non-invasive technique in the field of cognitive neuroscience that makes few physical demands of study participants. Although there are now several research techniques that can be used to connect characteristics of EEG recordings with cognition, only two of them have been used in theory of mind research. The first, and most common, is the event-related potential (ERP) technique in which the EEG is recorded time locked to the presentation of a particular stimulus that requires a theory of mind judgment. The EEG signals from each trial are averaged to capture the stable, reliable characteristics of the brain response, which are typically what are reported in ERP studies. The second is correlating characteristics of resting-state or baseline recordings of EEG with theory of mind performance. In this brief review, we will summarize the findings from a small, but growing number of studies that have used these methods to characterize the neural correlates of theory of mind.

Event-related potential studies of belief/desire reasoning

A primary strength of the ERP technique is that it has the potential to capture in fine temporal detail (typically 1–4-millisecond resolution), and reasonable spatial detail, the neurocognitive events that are associated with the processing of a particular class of stimulus. There are, however, two challenges of applying the ERP methodology to the study of theory of mind. The first is that it is not entirely possible to determine precisely when someone has made an inference about someone's mental state. Take, for example, the standard false belief task, which is common in both developmental research and in much of the research that I will review below. In a typical false belief task, one story protagonist leaves an object in one hiding place and then leaves the scene. While that character is gone, a second character takes the object out of the original hiding place and places it in an alternative hiding place. At this point in the story, the first story character has a false belief about the location of the object. Yet, it is difficult to pinpoint when someone hearing the story might infer or realize that the first character has that false belief. Indeed, such inferences might be made slowly as information is integrated across time. Because of this difficulty, the time-locked nature of ERP is not obviously well-suited to investigating the neurocognitive processes by which individuals develop inferences about others' beliefs.

A second problem with applying the ERP method to standard theory of mind tasks is that, as with most cognitive neuroscience techniques, ERP typically requires participants to endure at least 40 trials of a particular condition type to maximize the signal-averaging benefits. While solving

one typical false belief task might rely on theory of mind skills, it seems possible that after 40 or more, participants might fall back into a routine or pattern in which they are solving the task in ways other than reasoning about false beliefs.

Sabbagh & Taylor (2000) carried out the first attempt to address these issues and apply the ERP technique to reasoning about false beliefs. In their tasks, participants read scenarios in which an actor placed two objects in a scene. After the first actor leaves the scene, a second actor enters and moves one of the objects that the first actor originally set. Thus, the first actor has a true belief about the location of one of the objects, and a false belief about the other. Following the story, a test question about the first actor's beliefs was presented (one word at a time) such that participants did not know which object was going to be asked about until the final word of the sentence (e.g. "According to Chester, where is the [object]"). In this way, Sabbagh & Taylor (2000) argued that, although theory of mind relevant inferences may have been occurring throughout the study, they had to be reasoning about the first actor's beliefs at the moment when they knew what question they were answering. Furthermore, because they were unsure whether they would be asked about reality or belief, they could not rely on simple response strategies to correctly answer questions.

Along with 40 false belief trials, participants were also given 40 "false photograph" trials that involved the same scenarios except that instead of the story being about an actor who left the room while an object was moved, the stories were about actors who took pictures of a room before things were moved. The false belief test questions had the same format, and even involved the same objects as the false belief trials (e.g. According to the photo, where is the [object]"). This control condition is important as it has many of the same surface task demands of the false belief task, but does not require reasoning about mental states.

ERPs elicited by the final word of the test question in the false belief and false photograph trials were compared and showed a clear, focal dissociation in a slow-wave component of the ERP over left anterior frontal regions. Specifically, the slow wave associated with belief reasoning was more positive than the slow wave associated with photo reasoning. The dissociation emerged at around 300 milliseconds after the onset of the sentence final word of the test question and was maintained throughout the rest of the ERP recording epoch (1000 milliseconds). The findings also revealed a difference, emerging at roughly the same time, over a positive component at parietal sites whereby belief reasoning had a diminished amplitude relative to reasoning about photographs. No source localization analyses were performed, but the findings were generally consistent with cortical sources in the medial frontal regions, and the inferior parietal regions, both of which have shown sensitivity to the distinction between belief and photo reasoning in other neuroimaging studies (e.g. Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2005).

Another study (Liu, Sabbagh, Gehring & Wellman, 2005) used a similar methodology, with two key differences. The first difference was that instead of reading false belief stories, participants watched simplified cartoon-style animations of a scene in which a character was left with a true belief about the location of one object, and a false belief about the location of the other. The stories were narrated by an experimenter in real-time. At the end of the story, the narrator asked the test question ("Where will Garfield look for this") and then showed a picture of either the false belief or the true belief object. Thus, instead of a word being the stimulus that elicited the ERP, here it was the picture. The second change was the comparison condition—in this case, the ERPs elicited after the belief questions were contrasted with ERPs elicited after a question about reality (i.e. "Really, where is this?"). Although judgments about reality are arguably not as well matched with belief judgments as were the photo judgments from the previous study, the results of the two studies were remarkably similar. Specifically, the ERPs elicited in the belief and reality conditions were dissociated in a slow wave component over left anterior frontal areas, beginning at about 300 milliseconds

post-stimulus, just as in Sabbagh & Taylor (2000). The only difference was that whereas the belief ERP was more positive than the photo ERP in Sabbagh & Taylor (2000), here the belief ERP was more negative than the reality ERP. It is unclear what to make of this difference, particularly given the similarity of the timing and the scalp topography of the dissociation. Perhaps more surprising, this study did not provide strong evidence for a corresponding slow wave effect in parietal regions. Nonetheless, these two ERP studies are consistent in showing that frontal lobes make an early contribution to theory of mind reasoning.

The ERP findings regarding a late slow wave effect (either positive or negative) associated with reasoning about false beliefs has been replicated with many different kinds of stimuli, and also now across multiple labs in other countries (e.g. Germany: Meinhardt, Sodian, Thoermer, Dohnel, & Sommer, 2011; China: Wang, Liu, Gao, Chen, Zhang & Lin, 2008; Zhang, Sha, Zheng Ouyang, & Li, 2009). To gain further evidence regarding the contribution that the frontal slow wave might be making to false belief reasoning, David Liu and colleagues (Liu, Meltzoff & Wellman, 2009a) adapted tasks that have been used with preschool-aged children for measuring reasoning about different kinds of mental states—beliefs and desires—for use in an ERP paradigm. In the key trials, children were told about two story characters who had either different desires or different beliefs that were relevant to the contents of a closed box. The box was then opened, and the contents were revealed to be consistent with one character's beliefs or desires. ERPs were recorded time locked to the opening of the box, which were preceded by a question designed to guide participants' processing about beliefs or desires (e.g. "Who will say, 'I want some,' when they see this ...?" or "Who will say, 'I was wrong,' when they see this ...?"). ERPs elicited in these conditions were compared with those elicited in a control condition that asked about where the items that come in the box might be put away. Results showed that a positive frontal slow wave (similar to that shown in Sabbagh & Taylor, 2000) was stronger (relative to control) in the belief and desire conditions. This paradigm did elicit a right lateralized parietal slow wave effect, which was stronger (relative to control) in the belief condition only. These findings join recent findings from fMRI (e.g. Saxe & Powell, 2006) in suggesting that frontal lobe contributions to theory of mind might be important for representing mental states more generally, whereas parietal lobe contributions might be particularly important for reasoning about beliefs specifically.

Studies with children

Along with being relatively inexpensive, EEG methods have the general advantage of perhaps the least taxing of the cognitive neuroscience techniques from the standpoint of the participant. Even dense-array EEG (up to 256 channels) acquisition "hats" or "nets" can be applied comfortably in minutes by a single researcher in an open room. Because of this convenience, EEG/ERP methods can, at least in principle, be applied to participants from a wide range of ages. This is a particularly important advantage for theory of mind research, because much of the interest in this area has centered on the rapid, generally stereotyped development of theory of mind skills in preschool-aged children, and the neurobiological events that may contribute to abnormalities in its development during early childhood (as in the case of autism). Because the same recording methods and experimental paradigms can (again, at least in principle) be used with children of different ages and with adults, EEG/ERP methods provide a clear opportunity for looking at how the neural mechanisms that are associated with various aspects of theory of mind reasoning change over time.

This was recently attempted by Liu and colleagues (Liu, Sabbagh, Gehring & Wellman, 2009b) who used the same basic paradigm that this group had used in the study described above (Liu, Sabbagh, Gehring & Wellman, 2004) with a group of 6-year-old children and a second group

of adults. Unlike standard false belief tasks that 4-year-olds pass easily, the ERP false belief task required children to track and answer questions about two mental states (a true belief and a false belief) as opposed to just one. Accordingly, 6-year-olds varied in their behavioral performance on this false belief task and many showed systematically poor performance. This variability allowed for the comparison of ERP effects among “passers” and (for lack of a better term) “failers.” The findings here were striking. Adults showed the same pattern that was seen in Liu et al. (2005)—a slow wave dissociation emerging at right anterior frontal sites 300 milliseconds post-stimulus with the belief ERP being more negative than the reality ERP. For the children, false belief “passers” showed the same effect as the adults, though its timing was somewhat later and the scalp distribution was somewhat broader, likely reflecting that increased efficiency of neural processing with age and development (see e.g. Johnson, 2001). False belief “failers,” however, showed no systematic dissociation between belief and reality ERPs over left frontal areas, or any other electrode sites. These findings show that focal frontal lobe contributions are critical to theory of mind reasoning both in adults and young children.

Even more recently, Lindsay Bowman and colleagues (Bowman, Liu, Meltzoff & Wellman, 2012) adapted the paradigm for comparing the neural correlates of belief and desire reasoning for use with 8- and 9-year-old children. These findings showed that, like the adults, there was a left lateralized late slow wave effect associated with reasoning about both beliefs and desires. The other effect that was seen in adults, a right lateralized slow wave effect over parietal areas for reasoning about beliefs was seen when analyzing trials on which children showed accurate performance. These findings suggest that the functional recruitment of right parietal areas for reasoning specifically about beliefs may be a later developing feature of the neural correlates of theory of mind reasoning (see also Saxe et al., 2009).

Using a different paradigm, Meinhardt and colleagues (Meinhardt, et al., 2011) came to similar conclusions about both the consistency of the frontal late slow wave (LSW) in reasoning about false beliefs, and developmental changes in the contribution of the parietal regions. In their study, 6–8-year-old children and adults watched vignettes in which a story character developed either a true or a false belief about the location of an object. At the end of the vignette, ERPs were recorded as participants watched the character acting either in accordance with his beliefs (true or false) or unexpectedly (i.e. counter to their beliefs). Results showed that for children and adults reasoning about false beliefs was associated with a frontal LSW and parietal effect not seen for reasoning about true beliefs. For adults, the parietal effect had a somewhat more central distribution whereas for children, the parietal effect was focused (although still broadly distributed) over parietal regions.

Comparisons between the scalp distributions of these findings to those from other studies are limited by differences in the electrode referencing schemes that were used (for a fuller discussion, see “Issues and new directions for ERP studies”). Nonetheless, the findings do converge on a general developmental picture in which children’s early theory of mind reasoning may rely critically on neurocognitive processes within medial frontal regions, whereas parietal contributions to theory of mind reasoning that are seen in adults may continue to develop and be refined through late childhood.

ERP studies of mental state decoding

In the belief/desire reasoning tasks described above, participants are asked to use contextual background information about an individual to make an inference about that individual’s likely mental state, and then assess how that person might act in a given situation based upon that mental state. Another, complementary, aspect of theory of mind is what we might call “decoding” others’ mental

states—the process associated with determining others’ mental states based less on our idiosyncratic knowledge of other persons and their histories and more on immediately available information such as gaze direction, facial expression, speech prosody, and so forth (see e.g. Sabbagh, 2004). Of course, mental state decoding and belief/desire reasoning work in concert to render accurate judgments about others’ mental states in everyday situations. Nonetheless, because the two processes rely on fundamentally different kinds of information, they may also rely on fundamentally distinct neurocognitive processes.

In some ways, the ERP technique is highly amenable to the study of mental state decoding. First, it is relatively easy to control the moment that someone engages in mental state decoding. Secondly, because the process does not necessarily rely on developing a context of idiosyncratic personal histories the way belief/desire reasoning does, the necessary numbers of trials can be carried out quickly. Yet, despite this natural fit, we know of very little work that has used ERP to examine the neurocognitive processes associated with mental state decoding. Indeed, the only study we know about was done by Sabbagh, Moulson, & Harkness (2004), who adapted Baron-Cohen Wheelwright, Hill, Raste, & Plumb’s (2001) “Reading the Mind in the Eyes” task for use in an ERP paradigm. In their study, adults saw either a mental state term (e.g. “desiring”) or a sex term (e.g. “female”) that was followed by a picture of the eye region of a face. The participants’ task was to determine whether the mental state term or the sex term was an accurate description of eye picture. ERPs were recorded to the presentation of the picture of the eyes, as this was when participants were decoding either the mental states or the sex depicted in the picture. Results showed that there were two ERP effects that reliably dissociated trials in which participants made judgments about mental states from those in which participants made sex judgments. The first was an N270 effect over right anterior frontal regions, and the second was a right temporal negative slow wave effect. For both of these components, ERPs associated with mental state decoding were more negative than those associated with sex decoding.

Although Sabbagh et al. (2004) is the only study that has used ERP to investigate the processes of mental state decoding from within the theory of mind framework, there are two bodies of ERP literature that are highly relevant to this question. The first concerns the ERP correlates of gaze direction perception. Gaze direction perception is highly relevant to mental state decoding as it provides information about other’s attentional states (e.g. Baron-Cohen, 1994). Research from several laboratories has suggested that much of the brain’s specialized electrophysiological response to faces—specifically, the N170 component over occipital-temporal regions) may result from processing that is specific to the eye region of the face (e.g. Itier, Alain, Kvacevic & McIntosh, 2007; Puce, Allison, Bentin, Gore & McCarthy, 1998). However, this specialized component does not appear to reliably index sensitivity to gaze direction (see Itier & Batty, 2009 for a review). Instead, the cognitive operations that are associated with making judgments related to gaze direction appear later in the ERP record and have a more anterior temporal frontal distribution (e.g. Itier et al., 2007; Conty, N’Diaye, Tijus, & George, 2007). In one particularly interesting study, Senju and colleagues (Senju, Tojo, Yaguchi, & Hasegawa, 2005) compared the ERPs elicited in an oddball paradigm in which pictures that showed an actor either displaying direct or averted gaze were targets in the context of a standard that showed an actor with downcast eyes. Results showed that perception of gaze direction (both direct and averted) was associated with a right temporal N270 component, similar to that seen in the Sabbagh et al. (2004) mental state decoding paper. Considering these findings together, it may be that the temporal N270 component is associated with mental state processing associated with gaze direction processing. What was particularly notable about Senju et al.’s (2005) findings was that 12-year-olds with autism also showed the N270 effect when making judgments about gaze direction. However, the effect was bilaterally

distributed, thereby suggesting that the organization of neural systems for decoding mental states may differ in autism as compared with typically developing children.

A second literature that is directly relevant to mental state decoding concerns emotion recognition. Emotion recognition—typically operationalized as the ability to accurately identify other's emotions based upon facial expressions—can be thought of as a special case of mental state decoding insofar as both involve making a judgment about someone's mental state based upon available perceptual information. There are many ERP studies of facial emotion recognition, most of which investigate differences in neural responding to the six basic emotions (fear, anger, happiness, sadness, disgust, surprise), with a particular interest in understanding (1) whether the neural response to negative emotions differs from positive or neutral emotions, and (2) how the emotionality of the face affects the early perceptual processing of facial features and configurations, and 3) how these effects are modulated by attention (see e.g. Batty & Taylor, 2003; Eimer, Holmes & McGlone, 2003 for nice examples). Within this framework, however, little attention has been paid to the processes underlying the decoding of more subtle emotional expressions that are likely to be important in everyday mental state attribution (e.g. detecting that an interlocutor is confused). One recent attempt was undertaken by Debrulle, Brodeur, & Hess (2011) who presented participants with full face stimuli displaying ambiguous expressions (models were in fact not asked to display any particular expression). Participants were asked to judge whether a given face was neutral, positive, negative, or ambiguous. Results were broadly consistent with prior findings (including Sabbagh et al., 2004) in showing that (among other things) N270 component over anterior frontal regions discriminated neutral from valenced judgments. Although more work needs to be done to clarify the exact nature of these effects, it does appear that the N270 component of visually evoked ERPs may be a reliable index of the neurocognitive systems critical for mental state decoding.

Issues and new directions for ERP studies

In a sense, the studies that we have described so far share an experimental logic with PET or fMRI studies that have sought to “localize” the neural regions that subserve some aspect of theory of mind reasoning. That is, they have sought to compare the neural activations elicited by mental state reasoning with those elicited in some control condition (i.e. reality reasoning or photograph reasoning) with the aim of identifying special regions of the brain that are engaged for reasoning about mental states. However, the term “localize” does not readily apply to ERP effects. As is apparent in the above studies, the spatial distribution of a particular EEG/ERP effect can be broad, which on its own allows for only a general characterization about the location of the effect (e.g. inferior frontal region).

There are two problems that pose difficulties for more detailed characterization of sources. The first, and most serious, is what is sometimes called the “equivalent dipole problem”—any given ERP effect is mathematically consistent with more than one cortical source. Of course, this is not to say that some degree of reliable inference about cortical regions can be made for ERP effects. Indeed, methods for accurately localizing cortical sources of ERP effects are improving rapidly due to better recording techniques that allow for high spatial density sampling of scalp electrical fields, and increased understanding of the physical and anatomical constraints that govern how electrical signals propagate from neural dipoles to the scalp.

The second is a less obvious technical difference in how different researchers characterize EEG at a given electrode site. Voltage is measured as a difference between two sites—an active site and a “reference” site where there is supposed to be an approximation of zero activity. Some researchers assume a “mastoid reference” which is the average activity recorded at electrodes placed on the

mastoid bones, where there is thought to be little or no activity. Others using a dense recording array (up to 256 channels) use an “average reference” wherein the average of all electrodes (which, given the dipolar nature of neuroelectric generators, should theoretically be zero) is used as the reference. A discussion of the pros and cons of each method is beyond the scope of this chapter, but, the choice that a given researcher makes has important implications for the characterizing the precise spatial distribution of condition effects, making it different to compare across studies. Although somewhat arcane, the problem is quite serious. For instance, although frontal LSW effects are apparent in all studies that have sought to discriminate mental state (and especially false belief) reasoning from reasoning in control conditions, differences in recording techniques made it difficult to know whether the effects are homologous or heterogenous across studies.

Of course, these spatial imprecisions can be reduced by adopting sound practices in recording and analyzing EEG (i.e. use of many electrodes spaced evenly over the scalp, precise characterization of electrode positions using photogrammetry or 3D imaging, etc.). Also, while researchers should strive to minimize these imprecisions, some level might be tolerable given the low-cost and ease of EEG recording. As was noted above, EEG is readily applicable to a wide variety of sensitive populations (e.g. children, individuals with autism), and this advantage has made it a reliable first-pass technique for understanding the neurocognitive underpinnings of various cognitive processes. The fact that the extant studies on false belief reasoning and mental state understanding more generally have rendered such similar findings is certainly impressive and can be taken as further evidence that the same is true for the role of EEG in understanding the neurocognitive systems underlying theory of mind.

These studies that have used ERP within a “localization” approach, however, should not overshadow a more distinct advantage of ERP which is that it provides an exquisitely sensitive characterization of the timing of a particular neurocognitive process. This advantage may be particularly important for understanding the neural bases of theory of mind. Arguably, deploying theory of mind reasoning even in a highly constrained task context requires the coordination of multiple processes. For instance, to make a judgment about others’ mental states, one must generate a representation of the semantic content of those mental states (which itself can rely on numerous factors, such as remembering a person’s specific history, etc.) and enact whatever executive processes are necessary for decision making based upon those representations of others’ mental states. Theorizing about both how to best characterize these cognitive operations and their interplay is critically constrained by understanding the timing of their relative contributions to final theory of mind judgments.

In one such study, McCleery and colleagues (McCleery, Surtees, Graham, Richards & Apperly, 2011) adopted a task developed to examine the cognitive dynamics of visual perspective taking for use in an ERP paradigm. In this task, participants were shown a picture of a stage set in which three walls were visible (left, back, and right). On the walls of the room, there was some number of “disks” (black dots) in some arrangement and all visible to participants. The key manipulation was that also within the room there was a character (or “avatar”) who had either a full or partial view of the total disks. On some trials, the participants’ task was to say how many disks participants themselves see and on other trials say how many the avatar sees. A key manipulation was that some of the time, participants saw the same as the avatars, whereas other times there was a discrepancy. This design allows for main effect comparisons of (1) the neural mechanisms associated with perspective taking (through the self-other comparison) and (2) the neural mechanisms associated with resolving conflict between two response options (when participants and avatars see the same vs. different numbers of disks). Although some there were some complexities in the results, the findings were clear in showing that right (and to some extent, left) posterior regions were the first to index a difference in making judgments about one’s own vs. another’s perspective,

with differences emerging on a slow positive ERP component, similar to the posterior slow wave that was present in the ERP studies described above. This component was slower to peak when making judgments about others' perspective than when making judgments about one's own. Later in the ERP, a late slow wave 600–800 milliseconds post-stimulus, a lateral frontal component differentiated instances in which individuals had to make a judgment when participants and avatars saw the same vs. different numbers of disks. Integrating these findings with others from the ERP and broader neuroimaging literature, the authors concluded that a posterior temporal parietal system is associated with computing differences in visual perspective whereas the frontal system is associated with selecting the appropriate response from the conflicting options. Most compelling was that the timing of the effects showed that the computation of another's perspective preceded the response execution.

Still another advantage of ERP studies is that the components of an evoked potential are largely similar across a wide range of studies. For instance, ERPs that are evoked with visual stimuli (as has been the case with all studies to date) have a typical signature response that has been well characterized across numerous studies over time. Three such components that are relevant to understanding the neural processes associated with theory of mind are the N2 (a negative deflection over central-frontal areas peaking around 200 milliseconds post-stimulus), P3 a slower, positive deflection over parietal areas peaking between 300–600 milliseconds post-stimulus), and LSW, which can be negative or positive over lateral frontal regions that emerges around 500 milliseconds post-stimulus and continues to the end of a 1 second of the recording epoch. A long history of ERP research in a more psychophysiological tradition has aimed to catalog the types of manipulations that alter the timing, amplitudes, and spatial distributions of these effects. The result is that ERP effects that are elicited in theory of mind studies such as those described above can be constrained by these standard psychophysiological interpretations of the components. Some researchers have applied this kind of reasoning to their results. For instance, Liu et al. (2009b) leveraged a large ERP literature to marshal evidence that the left lateralized negative LSW effect for false belief reasoning they saw in children (who passed) and adults reflected ongoing conceptual operations in working memory required to reason through false belief scenarios. Similarly, McCleery, et al. (2011) relied on past work to argue that their centrally distributed positive LSW effects reflected the operations necessary for negotiating conflicting response options, which is a critical task demand in most theory of mind judgments. These constraints on theorizing provide a distinct advantage that complement other neuroimaging work in theory of mind.

More theoretical and methodological work can be done to take advantage of the psychophysiological aspects of ERPs to better understand the neurocognitive underpinnings of theory of mind. For instance, the parietal slow wave component that is typically associated with belief reasoning is very much akin to the later P3 component of the visual evoked potential (sometimes called the P3b). The P3 is one of the most discussed ERP components in the literature and even a partial review of this literature is beyond the scope of this chapter (but see Polich, 2007 for a recent review). A particularly promising direction is to develop theory of mind paradigms that can take advantage of the P3b and its theoretical interpretations to better understand the cognitive contribution that parietal regions make to theory of mind reasoning.

Resting state EEG studies

Studies with adults

During awake mental relaxation, EEG alpha (8–13 Hz) becomes synchronized and thus amplified across the scalp. In contrast, during mental activity, activity in the alpha band becomes

desynchronized as cortical circuits engage in task-specific cognitive activity (see Klimesch, 1999, for a review). Thus, EEG alpha power can serve as a reasonably precise inverse measure of cortical activity; that is, relatively stronger alpha power reflects less mental activity (Gevins, 1998). Within this context, EEG alpha has been used as a measure of individual differences in tonic cortical activation. In particular, researchers have been focused on how regional differences in tonic cortical activation may be a trait-like characteristic that might predispose individuals to particular cognitive or affective styles (Davidson, 1998; Hagemann, Naumann, Thayer, & Bartussek, 2002).

To the extent that there are individual differences in theory of mind skills, we might expect that different patterns of tonic cortical activation would provide insight into the neurobiological bases of these differences. To our knowledge, there has been only one such attempt. Sabbagh & Flynn (2006) explored whether individual differences in healthy university students' resting state EEG might be associated with mental state decoding. Using the reading the mind in the eyes task (as did Sabbagh et al., 2004) results from individual differences and group analyses showed that individuals with tonic activation at right mid-frontal leads was positively associated with performance on the mental state decoding task. There were no significant effects at posterior parietal leads. These findings suggest that stable, individual differences in mental state decoding might be associated with tonic activation of the right frontal regions.

Although there has only been one attempt to link characteristics of resting EEG to individual differences in theory of mind performance in adults, we think that there are a number of reasons that this could represent an important direction for future research. In adults and children, a large body of work in affective neuroscience has used measures of EEG activation to better understand individual differences in different aspects of affective style, which themselves are associated with social competence in everyday situations (see e.g. Davidson, 1998). This work might offer an opportunity to provide an unexpected link between affective style and theory of mind. For example, like mental state decoding, clinical mood disorder symptoms of anhedonia and dysphoria are also associated with increased tonic activation of right frontal regions (e.g. Gotlib, Ranganath & Rosenfeld, 1998). Interestingly, a growing body of literature suggests that dysphoric individuals are better at mental state decoding than non-dysphoric controls (e.g. Harkness, Sabbagh, Jacobson, Chowdrey, & Chen, 2005; Harkness, Jacobson, Duong, & Sabbagh, 2010). Further research with EEG will be useful in further solidifying the connection between affective style and theory of mind.

Studies with children

Resting EEG can be used in the developmental context to provide a reliable measure of functional cortical maturation (Thatcher, 1992). Over the preschool period, the alpha rhythm (for children between 6 and 9 Hz, as opposed to 8–12 Hz for adults) becomes increasingly dominant in the baseline EEG. Most important for the present purposes, there are regional changes in alpha coherence (the non-linear correlation in signal at any two electrodes). These changes in alpha coherence are caused by increases in synchronized neuronal firing both within and across neural populations that reflect the developmental changes in the organization of neurocognitive systems (Nunez, 1995). Recent advances in EEG analysis, such as standardized low-resolution electromagnetic tomography (sLORETA; Pascual-Marqui, 2002) have made it possible to use coherence measures to estimate the intracerebral sources of spectral EEG power. When applied in the developmental context, we can assume that regional changes in source-localized current density estimated from baseline EEG recordings reflect ongoing neurodevelopmental processes within that region.

With this in mind, Sabbagh, Bowman, Evraire, and Ito (2009) investigated the association between regional cortical maturation and performance on a battery of standard theory of mind tasks to better understand the neural bases of theory of mind in preschool-aged children. Of particular interest

was whether the neural systems whose development is positively associated with theory of mind development are homologous with those that are important for theory of mind reasoning in adults. The results from the source-localization analyses were clear in showing that individual differences in the current-source estimates attributable to the dorsal medial prefrontal cortex (dMPFC) and the right temporal parietal juncture (RTPJ) were associated with preschoolers' performance on the theory of mind battery. These associations were present when individual differences in preschoolers' executive functioning and language skills were statistically controlled. What is more, these regions showed substantial overlap with those that have been identified in fMRI studies of theory of mind reasoning in adults (e.g. Saxe, 2006).

Sabbagh et al. (2009) pointed out several limitations in their findings. Among the most important concerned the nature of the source-localization analyses. As noted above, cortical source localization based upon EEG is problematic because of the equivalent dipole problem, which affects these findings as well. Of course, the equivalent dipole problem is minimized as increasingly accurate models of how electrical signals propagate through the brain, skull, and scalp. For now, the average parameters for these models are well-characterized for adults, although, considerably less work has been done to establish the appropriate parameters for estimating sources in children. Because of this, Sabbagh, et al. (2009) used the adult models and urged caution in interpreting their results.

A second, more theoretical issue, is that there are fundamental ambiguities surrounding how to interpret this concurrent relation between regional current-source density and children's theory of mind skills. One possibility, favored by Sabbagh et al. (2009), is that the functional maturation of these regions may constitute a rate-limiting factor on the emergence of explicit theory of mind skills in preschool aged children such that as these regions become more functionally mature so too does children's ability to explicitly reason about representational mental states. Of course, both endogenous and exogenous factors might affect the maturation of these regions, but the critical point is that the maturation of these regions is itself important to children's theory of mind development. An alternative possibility is that somehow children's explicit representational theory of mind skills spur the functional maturation of these brain regions. The EEG measures that Sabbagh et al. (2009) used are well suited to using longitudinal research that can do a better job of establishing causal relations among these different developmental achievements.

Mu-suppression

Along with providing insight into the biological bases of young children's theory of mind development, a small group of researchers have been using EEG measures as a way of understanding infants' understanding of intentional action. Although alpha rhythms can be detected to some degree at all regions of an EEG recording array, there are some regional differences in character. One particularly well-studied characteristic of alpha is the so-called mu rhythm which is recorded from leads positioned over the motor cortex (e.g. Cz, C3, C4). The mu rhythm is desynchronized (or "suppressed") during intentional motor planning and execution, presumably because the neural circuits that were generating the mu rhythm at baseline (rest) become engaged to complete the motor task.

The phenomenon of mu suppression has recently been leveraged by researchers interested in "mirror neuron" approaches to the origins of social cognitive processes. A detailed review of the general approach is beyond the scope of this chapter. In brief, the term "mirroring" is used to capture instances in which a particular neural mechanism becomes active in both the production and observation of intentional actions (Jeannerod, 2001; Prinz, 1997). Mirroring is of particular interest to researchers in social cognition because the mechanism seems to provide a framework

for interpreting others' goal-oriented actions—seeing one's own actions as equivalent to another's elevates the hypothesis that others' intentions are similar to one's own in the same context.

In one study, Southgate, Johnson, El Karoui & Csibra (2010) used the phenomenon of mu suppression to investigate the extent to which 9-month-old infants are able to predict others' goals in action. In their study, infants were presented with a hand posed in a "grasp" shape about to reach behind an occluder, which to an adult looks like the beginning of an attempt to grab a hidden object. In three control conditions, the same stimulus elements were maintained but reconfigured so that a detection of an impending "grasp object" intention was less likely. Results showed that the mu rhythm was desynchronized only in the stimulus condition in which it appeared that the hand was about to grasp something behind an occluder. These findings are striking in that the infants never saw the action completed—thus, the authors argue that they perceived the action as goal-directed and in doing so, activated the neural circuitry that is associated with the observation, planning and execution of goal-directed action. These findings were also consistent with an earlier study that showed similar findings of mu-suppression with children observing the actual outcomes of goal-directed activities (Southgate, Johnson, Osborne & Csibra, 2009).

While these findings are intriguing, some caution should be used in their interpretation. Although data were recorded from all over the scalp, only the data from the central electrodes where mu is typically found are reported. It is possible that alpha suppression occurred over the scalp more generally during action observation, which may raise questions about whether the true nature of the motor cortex's involvement. One possibility is that the suppression seen during action observation comes from a source other than motor cortex. Some hint that this may be true comes from Marshall, Young & Meltzoff (2011) who recently sought to specify in more detail the scalp topography of mu suppression during action observation and execution with the goal of determining whether mu suppression can be considered as a neurophysiological correlate of the human mirror system. In their study, infants were presented with a live actor who demonstrated a simple novel action—pressing a button on a box to make a sound. Then, infants alternated between watching the experimenter perform the action, or performing the action themselves, while EEG was continuously recorded. Results showed that when infants performed the action themselves, the mu rhythm was suppressed over central sites (located over the motor cortex) and not elsewhere on the scalp. In contrast, when infants watched others perform the action, there was evidence of alpha attenuation not just at the central sites but all across the scalp. The broad distribution of alpha suppression more generally at all sites during observation may suggest one of two things. Either action observation activates a large cortical network within which it is unclear whether the central sites are representing an independent contribution to that process. Or, the broad topography of the alpha attenuation indicates a broader, potentially deeper cortical process that spreads effects to the central recording sites. Clearly, more work is necessary to confirm the extent to which mu-suppression can be seen as a neural correlate of the mirror system that is arguably important for theory of mind development.

Conclusions

Here, I have reviewed how brain electrophysiological recordings, ERP/EEG, have helped to better understand the neural and neurodevelopmental bases of theory of mind, and outlined ways in which their continued use will provide an important complement to other neuroimaging methods. It is our hope that given its ease of use and, hopefully, increasingly standardized methods of EEG recording and analysis, electrophysiological data of the kind described above can provide critical data particularly with respect to the neural correlates of the development of theory of mind, its precursors, and its abnormalities in populations with known neurogenetic disorders.

References

- Baron-Cohen, S. (1994). How to build a baby that can read minds. *Current Psychology of Cognition* **13**, 513–52.
- 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**, 241–51.
- Batty, M., & Taylor, M.J. (2003). Early processing of the six basic facial emotional expressions. *Cognitive Brain Research* **17**, 613–20.
- Bowman, L., Liu, D., Meltzoff, A. & Wellman, H.M. (2012). Neural correlates of belief and desire reasoning in 7- and 8-year-old children: An event-related potential study. *Developmental Science* **15**, 618–32.
- Conty, L., N'Diaye, K., Tijus, C., & George, N. (2007). When eye creates the contact! ERP evidence for early dissociation between direct and averted gaze motion processing. *Neuropsychologia* **45**, 3024–3037.
- Davidson, R. J. (1998). Anterior electrophysiological asymmetries, emotion, and depression: Conceptual and methodological conundrums. *Psychophysiology* **35**, 607–14.
- Debruille, B., Brodeur, M., B., & Hess, U. (2011). Assessing the way people look to judge their intentions. *Emotion* **11**, 533–43.
- Eimer, M., Holmes, A., & McGlone, F. (2003). The role of spatial attention in the processing of facial expression: An ERP study of rapid brain responses to six basic emotions. *Cognitive, Affective and Behavioral Neuroscience* **3**, 97–110.
- Gevins, A. (1998). The future of electroencephalography in assessing neurocognitive functioning. *Electroencephalography and Clinical Neurophysiology* **106**, 165–72.
- Gotlib, I.H., Ranganath, C. & Rosenfeld, J.P. (1998). Frontal EEG alpha asymmetry, depression and cognitive functioning. *Cognition and Emotion* **12**, 449–78.
- Hagemann, D., Naumann, E., Thayer, J.F., & Bartussek, D. (2002). Does resting electroencephalograph asymmetry reflect a trait? An application of latent state-trait theory. *Journal of Personality and Social Psychology* **82**, 619–41.
- Harkness, K.L., Jacobson, J.A., Duong, D. & Sabbagh, M.A. (2010). Mental state decoding in remitted major depression: Effects of sad versus happy mood induction. *Cognition and Emotion* **24**, 497–513
- Harkness, K.L., Sabbagh, M.A., Jacobson, J.A., Chowdrey, N., & Chen, T. (2005). Sensitivity to subtle social information in dysphoric college students: Evidence for an enhanced theory of mind. *Cognition and Emotion* **19**, 999–1026.
- Itier, R.J., Alain, C., Kovacevic, N., & McIntosh, A.R. (2007). Explicit vs. implicit gaze processing assessed by ERPs. *Brain Research* **1177**, 79–89.
- Itier, R.J., & Batty, M. (2009). Neural bases of eye and gaze processing: The core of social cognition. *Neuroscience & Biobehavioral Reviews* **33**, 843–63.
- Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage* **14**, S103–9.
- Johnson, M.H. (2001). Functional brain development in humans. *Nature Reviews Neuroscience* **2**, 490–501.
- Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews* **29**, 169–95.
- Liu, D., Meltzoff, A.N., & Wellman, H.M. (2009a). Neural correlates of belief- and desire-reasoning. *Child Development* **80**, 1163–71.
- Liu, D., Sabbagh, M.A., Gehring, W.J., & Wellman, H.M. (2004). Decoupling beliefs from reality in the brain: An ERP study of theory of mind. *NeuroReport* **15**, 991–5.
- Liu, D., Sabbagh, M.A., Gehring, W.J., & Wellman, H.M. (2005). An ERP study of 5-year-olds theory of mind. *Journal of Cognitive Neuroscience* **17**, 190–190.
- Liu, D., Sabbagh, M.A., Gehring, W.J., & Wellman, H.M. (2009b). Neural correlates of theory of mind reasoning in adults and children. *Child Development* **80**, 318–26.

- Marshall, P.J., Young, T., & Meltzoff, A.N. (2010). Neural correlates of action observation and execution in 14-month-old infants: An event-related EEG desynchronization study. *Developmental Science* 14, 474–80.
- McCleery, J.P., Surtees, A.D.R., Graham, K.A., Richards, J.E., & Apperly, I.A. (2011). The neural and cognitive time course of theory of mind. *Journal of Neuroscience* 31, 12849–54.
- Meinhardt, J., Sodian, B., Thoermer, C., Dohnel, K., & Sommer, M. (2011). True and false belief reasoning in children and adults: An event related potential study of theory of mind. *Developmental Cognitive Neuroscience* 1, 67–76.
- Millett, D. (2001). Hans Berger—From psychic energy to the EEG. *Perspectives in biology and medicine* 44, 522–42.
- Nunez, P.L. (1995). *Neocortical Dynamics and Human EEG Rhythms*. New York: Oxford University Press.
- Pascual-Marqui, R.D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): Technical details. *Methods and Findings in Experimental and Clinical Pharmacology* 24, 5–12.
- Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology* 118, 2128–48.
- Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology* 9, 129–54.
- Puce, A., Allison, T., Bentin, S., Gore, S.C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience* 18, 2188–99.
- Sabbagh, M.A. (2004). Understanding orbitofrontal contributions to theory-of-mind reasoning: Implications for autism. *Brain and Cognition* 55, 209–19.
- Sabbagh, M.A., Bowman, L.C., & Evraire, L.E. (2009). Neurodevelopmental correlates of theory of mind in preschool children. *Child Development* 80, 1147–62.
- Sabbagh, M.A., & Flynn, J. (2006). Mid-frontal EEG alpha asymmetries predict individual differences in one aspect of theory of mind: Mental state decoding. *Social Neuroscience* 1, 299–308.
- Sabbagh, M.A., Moulson, M.C., & Harkness, K.L. (2004). Neural correlates of mental state decoding in human adults: An event-related potential study. *Journal of Cognitive Neuroscience* 16, 415–26.
- Sabbagh, M.A., & Taylor, M. (2000). Neural correlates of theory-of-mind reasoning: An event-related potential study. *Psychological Science* 11, 46–50.
- Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology* 16, 235–9.
- Saxe, R., & Powell, L.J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science* 17, 692–9.
- Saxe, R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K.A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Development* 80, 1197–209.
- Senju, A., Tojo, Y., Yaguchi, K., & Hasegawa, T. (2005). Deviant gaze processing in children with autism: An ERP study. *Neuropsychologia* 43, 1297–306.
- Southgate, V., Johnson, M.H., El Karoui, I., & Csibra, G. (2010). Motor system activation reveals infants on-line prediction of others' goals. *Psychological Science* 21, 355–9.
- Southgate, V., Johnson, M.H., Osborne, T., & Csibra, G. (2009). Predictive motor activation during action observation in human infants. *Biology Letters* 5, 769–72.
- Thatcher, R.W. (1992). Cyclic cortical reorganization during early-childhood. *Brain and Cognition* 20, 24–50.
- Wang, Y.W., Liu, Y., Gao, Y.X., Chen, J., Zhang, W.X., & Lin, C.D. (2008). False belief reasoning in the brain: An ERP study. *Science in China. Series C, Life Sciences* 51, 72–9.
- Zhang, T., Sha, W.J., Zheng, X.R., Ouyang, H.L., & Li, H. (2009). Inhibiting one's own knowledge in false belief reasoning: An ERP study. *Neuroscience Letters* 467, 194–8.

Functional neuroimaging of theory of mind

Jorie Koster-Hale and Rebecca Saxe

Introduction

In the decade since the last edition of *Understanding Other Minds*, the number of papers that use human neuroimaging tools to investigate the neural basis of theory of mind (ToM) has exploded from four (described in Frith & Frith's 2000 chapter) to, as of 2013, well over 400. Studying ToM with neuroimaging works. Unlike many aspects of higher-level cognition, which tend to produce small and highly variable patterns of responses across individuals and tasks, ToM tasks generally elicit activity in an astonishingly robust and reliable group of brain regions. In fact, convergence on this answer came almost immediately. By 2000, Frith and Frith concluded that "studies in which volunteers have to make inferences about the mental states of others activate a number of brain areas, most notable the medial [pre]frontal cortex [(mPFC)] and temporo-parietal junction [(TPJ)]." These regions remain the focus of most neuroimaging studies of ToM and social cognition, more than a decade later (see Adolphs, 2009, 2010; Carrington & Bailey, 2009; Frith & Frith, 2012; and Van Overwalle, 2008 for some recent reviews). To our minds, this consensus is one of the most remarkable scientific contributions of human neuroimaging, and the one least foreshadowed by a century of animal neuroscience.

Nevertheless, most of the fundamental questions about **how** our brains allow us to understand other minds remain unanswered; we have mainly discovered where to look next. We hope that this gap means the next decade of neuroimaging ToM will be even more exciting than the last one. In this chapter, we offer a perspective on the contribution that neuroimaging has made to the science of ToM in the last decade, and some thoughts on the contribution that it could make in the next. The chapter has three sections: "Theory of mind and the brain" reviews the existing evidence for a basic association between thinking about people's thoughts and feelings and activity in this group of brain regions. "A strong hypothesis" discusses some objections, both theoretical and empirical, to a strong interpretation of this association, and our responses. "Where next?" highlights newer approaches to functional imaging data, which we expect will contribute to the future neuroscience of ToM, their strengths and their limitations.

Theory of mind and the brain

Theory of mind brain regions

Over the course of development, human children make a remarkable discovery: other people have minds both similar to and distinct from their own. Other people see the world from a different angle, have different desires and preferences, and acquire different knowledge and beliefs. Children learn that other people's minds contain representations of the world which are often true and

reasonable, but which may be strange, incomplete, or even entirely false. These discoveries (i.e. “building a Theory of Mind”) help children to make sense of some otherwise mystifying behaviors: why mom would eat broccoli, even though there is chocolate cake available (e.g. Repacholi & Gopnik 1997), or why she is looking for the milk in the fridge, even though dad just put it on the table (e.g. Wimmer & Perner 1983).

As readers of this volume know well, developmental psychologists historically focused on one key transition in this developmental process—when and how children come to understand false beliefs. Assessing understanding of false beliefs has been taken to be a good measure of ToM capacity because it requires a child to understand both that someone can maintain a representation of the world, and that this representation may not match the true state of reality or the child’s own beliefs. In a standard version of the false belief task, children might see that, while their mother thinks the milk is in the fridge (having put it there 5 minutes ago), it is now actually on the table. The children are asked: “Where will she look for the milk?” or “Why is she looking in the fridge?”. Five-year-old children, like adults, usually predict that she will look in the fridge, because that is where she thinks the milk is (Wellman, Cross, & Watson, 2001). Three-year-olds, however, predict that she will look on the table, explaining that she wants the milk and the milk is on the table (at least when asked explicitly; see, e.g. Onishi & Baillargeon (2005), Saxe (in press), and Southgate, Senju, & Csibra (2007), for further discussion of ToM behavior in pre-verbal children). In fact, when three-year-olds see her look in the fridge instead, some will go so far as to fabricate belief-independent explanations, stating that she no longer wants the milk, and must be looking for something else (Wellman et al., 2001; Wimmer & Perner 1983).

Building off decades of experience in developmental psychology, the first neuroimaging studies of ToM also used versions of false belief tasks. Adults, lying in positron emission tomography (PET) or magnetic resonance imaging (MRI) scanners, read short stories describing a person’s action (see Figure 9.2), and were asked to explain that action (usually silently to themselves, to avoid motion artifacts). These early studies revealed increased levels of blood oxygen and glucose uptake (indirect measures of metabolic activity, henceforth called “activity”), in a small, but consistent group of brain regions—left and right TPJ and mPFC, as noted by the Friths, and also medial parietal cortex (precuneus, PC) and more anterior regions of the superior temporal sulcus (STS), down to the temporal poles (Figure 9.1).

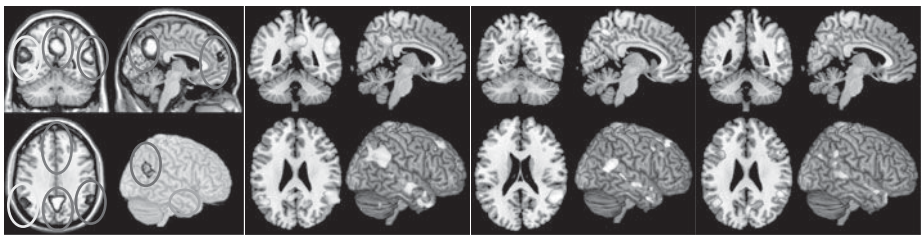


Figure 9.1 Brain regions commonly recruited in Theory of Mind tasks. (Left) Average activity in 63 subjects reading stories about false beliefs, compared to stories about false photographs ($P < 0.05$, corrected; see Dodell-Feder et al., 2011; Saxe & Kanwisher, 2003) overlaid on an average brain. Colored ellipses indicate standard locations of the right TPJ (red), left TPJ (yellow), right anterior STS (orange) medial parietal/precuneus (blue), and mPFC (green). Other three panels: activity in the same task in three example individual participants, overlaid on the same average brain anatomy for easy comparison ($P < 0.001$, uncorrected). Thanks to Nicholas Dufour for the images. See also Plate 2.

Activity during the false belief task is, of course, far from sufficient evidence that these brain regions have any role in understanding other minds. We believe this proposition becomes more compelling after reviewing the range of different experimental tasks and paradigms that have been used successfully over the years, across laboratories and countries, each aiming to elicit some aspect of “understanding other minds.” While some studies used complex verbal narratives, other researchers have used simple sentences or non-verbal cartoons; some studies explicitly instruct participants to think about a person’s thoughts, and others have elicited ToM spontaneously. The heterogeneity of methods, materials, and participant demographics makes it especially impressive that these studies have converged on the same regions of the brain. In this section, we will review a sample of these different procedures.

In the original theory of mind functional MRI experiments, both the content of the materials and the explicit instructions focused participants’ attention on (or away from) thinking about someone’s mind. For example, in an early PET study, Fletcher Happe, Frith, Baker, Dolan, Frackowiak, et al. (1995) told participants that they would be reading different kinds of verbal passages. Just before each item, the participant was told what kind of story was coming next. If it was a “mental” story, participants were instructed that it was “vital to consider the thoughts and feelings of the characters,” and then shown a story revolving around someone’s mental states (see example in Figure 9.2). If the story was a “physical” story, participants were first instructed that thinking about thoughts and feelings was irrelevant and undesirable, then shown a control story (see example in Figure 9.2). After each story, participants were instructed to silently answer an action-explanation question, such as “Why did the prisoner say that?”. Glucose consumption increased in the theory of mind brain regions while people read the mental stories, relative to the control stories.

The same design has been used with non-verbal stimuli. In an early fMRI experiment (Gallagher, Happé, Brunswick, Fletcher, Frith, & Frith 2000), participants both read the stories used by Fletcher et al. (1995) and were shown cartoons depicting visual jokes that either relied on ToM (in which understanding the joke depended on attribution of either a false belief or ignorance), or other types of humor (such as puns, idioms, and physical humor). Again, participants were cued in advance about whether to expect a mental or control cartoon (for examples, see Figure 9.2). For both types of cartoons, they were asked to silently contemplate the meaning; for mental cartoons, they were also explicitly instructed to consider the thoughts and feelings of the characters. With less than 20 minutes of scanning for each task, Gallagher and colleagues found that the same group of brain regions showed increased activity for both verbal and non-verbal ToM stimuli; these regions include the bilateral temporal-parietal junction and the middle prefrontal cortex. Similar convergence of the activity elicited by verbal and non-verbal stimuli has been found by Kobayashi, Glover & Temple (2007).

Sommer, Döhnelt, Sodian, Meinhardt, Thoermer, & Hajak (2007) also used non-verbal stimuli to focus participants’ attention on the thoughts of a character, but without explicitly cuing the condition. They showed participants a series of cartoon images depicting a story—Betty hides her ball, Nick moves it, and then Betty comes back to look for her ball. In half of the trials, Betty looks into the box where she thinks the ball is (the expected condition); in the other half, she looks into the other box (the unexpected condition). Participants judged whether, based on the character’s beliefs, the character’s action was expected or unexpected. Rather than explicitly labeling the mental and control conditions, a key contrast was between the beginning of the trial, before mental state inferences were possible, and the end of the trial, when participants had presumably made belief inferences to complete the task. The second key contrast was between trials in which Betty had a false belief (so that predicting her action required considering her thoughts) and trials in


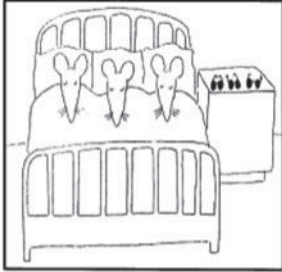
Mental	Control
<p>During the war, the Red army capture a member of the Blue army. They want him to tell them where his armies' tanks are. They know that they are either by the sea or in the mountains. They know that the prisoner will not want to tell them, he will want to save his army, and so he will certainly lie to them. The prisoner is very brave and very clever, he will not let them find his tanks. The tanks are really in the mountains. Now when the other side ask him where his tanks are he says, "They are in the mountains."</p>	<p>Two enemy powers have been at war for a very long time. Each army has won several battles, but now the outcome could go either way. The forces are equally matched. However, the Blue army is stronger than the Yellow army in air foot soldiers and artillery. But the yellow army is stronger than the Blue army in air power. On the day of the final battle, which will decide the outcome of the war, there is heavy fog over the mountains where the fighting is about to occur. Low-lying clouds hang over the soldiers. By the end of the day, the Blue army has won.</p>
	
<p>Brad had no money, but just had to have the beautiful ruby ring for his wife. Seeing no salespeople around, he quietly made his way closer to the counter. He was seen running out the door.</p>	<p>While playing in the waves, Sarah's Frisbee went flying toward the rocks in the shallow water. While searching for it, she stepped on a piece of glass. Sarah had to wear a bandage on her foot for a week.</p>
<p>The path to the castle leads via the lake. But children tell the tourists: "The way to the castle goes through the woods." The tourists now think that the castle is via the woods or lake?</p>	<p>The sign to the monastery points to the path through the woods. While playing the children make the sign point to the golf course. According to the sign the monastery is now in the direction of the golf course or woods?</p>
<p>How likely is Queen Elizabeth to think that keeping a diary is important?</p>	<p>How likely is Queen Elizabeth to sneeze when a cat is nearby?</p>
<p>John was on a hike with his girlfriend. He had an engagement ring in his pocket and at a beautiful overlook he proposed marriage. His girlfriend said that she could not marry him and began crying. John sat on a rock and looked at the ring.</p>	<p>Joe was playing soccer with his friends. He slid in to steal the ball away, but his cleat stuck in the grass and he rolled over his ankle, breaking his ankle and tearing the ligaments. His face was flushed as he rolled over.</p>
<p>That morning, people sat around looking at each other, wondering if they were dreaming, because everything looked purple. Some people were shocked. Some people thought that it was funny to see everybody all purple. But even the smartest scientists didn't know what had happened.</p>	<p>The whole world had turned purple overnight. Just about everything was purple, included the sky and the ocean and the mountains and the trees. The tallest skyscrapers and the tiniest ants were all purple. The bicycles and furniture and food were purple. Even the candy was purple.</p>
<p>In spite of her neighbourhood, Erica has a strong dislike of violence, and believes that conflicts can usually be resolved without fists.</p>	<p>Erica lives in Los Angeles. One night recently she was in a bar where a fight broke out between two drunk men and she was caught in between.</p>
<p>Sam thinks he can grow trees with fruit that taste like pizza. How likely is it that Sam wants these trees for a treehouse too?</p>	<p>In the backyard are trees with fruit that taste like pizza when ripe. How likely is it that these trees can be used for building a treehouse?</p>

Figure 9.2 Samples of experiments that elicit thinking about thoughts and feelings by manipulating the content of the stimuli. Sample stimuli from Fletcher et al. (1995), Gallagher et al. (2000), Mason & Just (2010), Perner et al. (2006), Lombardo et al. (2010), Bruneau et al. (2012), Saxe et al. (2009), Saxe & Wexler (2005), Young et al. (2010b).

which Betty knew where the ball was all along (so her action could be predicted based on the actual location of the ball).¹ Both contrasts revealed activity in ToM brain regions.

Another way to endow non-verbal stimuli with mentalistic content is by altering the movements of simple geometric shapes (Heider & Simmel, 1944). For example, in a PET study, Castelli, Happé, Frith, & Frith (2000) showed participants animations of two triangles moving around. Participants were instructed that while some of the triangles “just move about with random movement [...] disconnected from each other” (the control condition), other animations would show “two triangles doing something more complex together, as if they are taking into account their reciprocal feelings and thoughts [...] for example, courting each other” (the mental condition). Participants watched the animations, and then described what the triangles were doing. ToM animations elicited more activity than the random animations in the TPJ, the nearby superior temporal sulcus (STS), and the mPFC.

Some experiments elicit thinking about thoughts simply by describing those thoughts in words. For example, participants can answer questions about people’s mental characteristics, such as “How likely is Queen Elizabeth to think that keeping a diary is important?” vs. their physical traits, such as “How likely is Queen Elizabeth to sneeze when a cat is nearby?” (Lombardo, Chakrabarti, Bullmore, Wheelwright, Sadek, Suckling, et al., 2010; Mitchell, Macrae, & Banaji, 2006); or they can read single sentences describing thoughts (“He thinks that the nuts are rancid”) or facts (“It is likely that the nuts are rancid”; Zaitchik, Walker, Miller, LaViolette, Feczko, & Dickerson, 2010). In both cases, the items related to mental states elicited more activity in ToM regions than the control conditions.

Other experiments elicit thinking about thoughts indirectly. Saxe & Kanwisher (2003) used verbal stories based on either inferences about false beliefs or about physical events, similar to Fletcher et al. (1995). However, participants were not given any explicit instructions about the different kinds of stories. In Experiment 1, participants did not give any response, while in Experiment 2, they responded to fill-in-the-blank questions about details in the stories. Similarly, Mason & Just (2010) had participants read short stories about actions, and then answer simple comprehension questions. Critically, the stories elicited spontaneous inferences about unstated, but implied, events; some of these inferences were about a character’s thoughts (mental), and some about purely physical events (control). Compared with the original Fletcher et al. (1995) stories, the stories in these experiments (see examples in Figure 9.2) were shorter, and included less (or no) explicit description of thoughts and feelings (see also Bruneau, Pluta, & Saxe, 2011). Instead, the thoughts and feelings of the characters had to be inferred. Listening to the mental stories elicited strong activation in ToM regions relative to the control stories, suggesting consideration of the character’s thoughts despite the absence of explicit instruction.

Another procedure for eliciting spontaneous ToM in the scanner was developed by Spiers & Maguire (2006). Participants engaged in naturalistic actions (e.g. driving a taxicab through bustling London streets) in a rich virtual reality environment. After the scan and without prior warning, participants reviewed their performance, and were asked to recall their spontaneous thoughts

¹ In the original study, this contrast could have been due to a difference between false vs. true beliefs, or between representing a belief (required for the false trials) and making a prediction based solely on the actual location of the ball (possible for the true beliefs). Subsequent work has shown that activation observed by Sommer et al. (2007) is due to the latter: individuals often simply reduce the information they need to process by choosing not to represent true beliefs as a mental state (Apperly et al., 2007), and when this is controlled for, neuroimaging has revealed indistinguishably high activation for true and false beliefs (Döhnell, Schuwerk, Meinhardt, Sodian, Hajak, & Sommer, 2012; Jenkins & Mitchell, 2010; Young et al., 2010b).

during each of the events. These recollections were coded for content concerning the thoughts and intentions of the taxicab customers and the other drivers and pedestrians on the road (e.g. “I reckon that she’s going to change her mind”) and used these coded events to predict neural responses. They found that when participants were thinking about someone else’s intentions, but not during other events, regions in the ToM network showed increased activity.

Other studies have targeted spontaneous consideration of others’ intentions by asking participants to make moral judgments. Morally relevant facts appear to rely in part on consideration of intentions (Cushman, 2008), and evoke increased activity in the ToM network, relative to other facts in a story (Young & Saxe, 2009a). The same brain regions are recruited when participants are forced to choose between acting on a personal desire vs. a conflicting moral principle, compared to deciding between two conflicting personal desires (Sommer, Rothmayr, Döhnel, Meinhardt, Schwerdtner, Sodian, et al., 2010). These regions are also recruited in children watching an animation of one person intentionally harming another, compared to animations of other painful and non-painful situations (Decety, Michalska, & Akitsuki, 2008).

In fact, when participants read a story, they appear to automatically represent the thoughts and feelings of the characters in order to make sense of the plot, even if instructed to perform an orthogonal task. For example, Koster-Hale & Saxe (2011) had participants read short verbal stories, and then make a delayed-match-to-sample judgment, indicating whether a single probe word occurred in the story (match) or not (non-match); half of the stories described a false belief and half described physical representations. Despite the word-level task (and no mention of ToM in the explicit instructions), the contrast revealed activation across the ToM network. Similarly, in two fMRI experiments with children aged 5–12 years, children heard child-friendly verbal stories, describing characters’ thoughts and feelings vs. physical events; children answered orthogonal (delayed-match-to-sample) questions about each story. As with adults, we found increased activation in the ToM network when children were listening to stories involving thoughts and feelings (Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009; Gweon, Dodell-Feder, Bedny, & Saxe, 2012).

Thinking about thoughts and feelings can be also manipulated by changing the task while holding the stimuli constant (Figure 9.3). In an early PET study, Goel, Grafman, Sadato, & Hallett (1995) showed participants sets of 75 photographs of objects, some modern and familiar, and some from pre-fifteenth century North American aboriginal culture. Participants either judged whether the object was elongated along the principal axis (the control task) or whether “someone with the background knowledge of Christopher Columbus could infer the [object’s] function” (the mental task). They found increased ToM activation when participants were considering Christopher Columbus, but not when making the physical judgments. Similarly, Walter and colleagues showed participants sequences of three cartoon images (Schnell, Bluschke, Konradt, & Walter, 2011; Walter, Schnell, Erk, Arnold, Kirsch, Esslinger, et al., 2010). Participants either judged, on each picture, whether “the protagonist feels worse/equal/better, compared to the previous picture” (the mental task) or whether “the number of living beings [in the image is] smaller/equal/greater, compared to the previous picture” (the control task). In another set of studies, Baron-Cohen and others (Adams, Rule, Franklin Jr, Wang, Stevenson, Yoshikawa, 2010; Baron-Cohen & O’Riordan, 1999; Baron-Cohen, Wheelwright, & Hill, 2001; Platek, Keenan, Gallup, & Mohamed, 2004) showed participants pictures of a person’s eyes. Participants pressed a button to indicate either the mental/emotional state of the person in the picture (e.g. *embarrassed*, *flirtatious*, *worried*; the mental task) or their gender (the control task). Mitchell, Banaji, & MacRae (2005) asked participants to either judge either how happy a person was to be photographed (mental) or how symmetric their face was (control). In all of these cases, the mental tasks activated the ToM regions more than the control tasks.


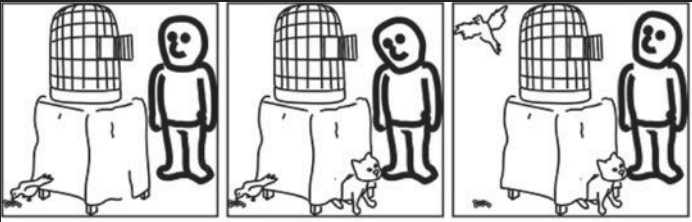

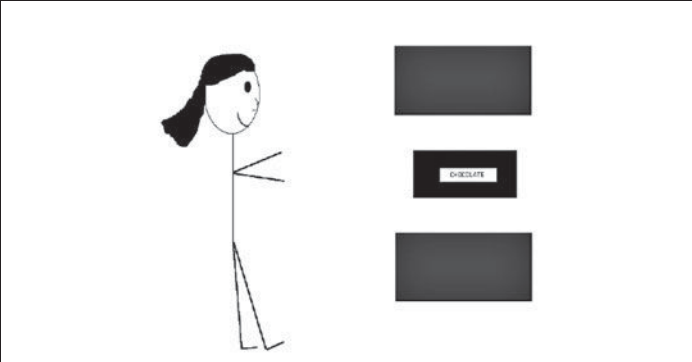
Stimulus	Mental Task
	Worried or friendly?
	(In each panel) Does he feel better or worse than in the previous panel?
	Why does she feel this way?
	Where does the girl think the chocolate is?

Figure 9.3 Samples of experiments that elicit thinking about thoughts and feelings by holding the stimulus constant, and manipulating the participants’ task. Examples from Adams et al. (2010), Schnell et al. (2010), Spunt & Lieberman (2012), and Saxe, Schulz & Jiang (2006).

Spunt and colleagues have recently developed a clever paradigm for eliciting ToM using a simple task manipulation. In their first experiment, Spunt, Satpute, & Lieberman (2011) showed participants pictures of simple human actions (e.g. a person riding a bike), and instructed them to silently answer one of three questions: why the person is doing the action (e.g. to get exercise), what the person is doing (e.g. riding a bike), or how the person is doing it (e.g. holding handle-bars). These questions require successively less consideration of the mind of the person and, correspondingly, showed successively less ToM region activity. Spunt & Lieberman (2012) replicated the result using a similar paradigm with brief naturalistic film clips of facial expressions of emotions. Participants judged either how the person is expressing their emotion (e.g. “looking down and away,” the control task) or why she is feeling that emotion (e.g. “she is confused because a friend let her down,” the mental task). Again, ToM regions were recruited more when thinking about why than how.

Using these types of parametric designs and analyzing the continuous magnitude of response in ToM regions may provide a powerful tool for studying the neural basis of ToM, especially in combination with computational models of ToM, which offer quantitative predictions of both when and how much (or how likely) people are thinking about others’ thoughts. For example, Bhatt, Lohrenz, Camerer, & Montague (2010) created a competitive buying and selling game in which participants could try to bluff about the value of an object. The authors predicted that reliance on ToM would increase in proportion to the *riskiness* of the bluff, i.e. the discrepancy between the object’s true value and the proposed price. Consistent with this idea, a region near the right TPJ showed activity correlated with bluff riskiness across trials. They suggested that participants may be more likely to engage in ToM, engage in more ToM, or engage in ToM for longer, when they are making a riskier bluff relative to a less risky bluff, and this relationship is continuous over a large range of possible risks.

Along with manipulating thinking about thoughts across stimuli and tasks, it is also possible to look at *when* participants are thinking about thoughts within a single ongoing stimulus and task. Stories about human actions and beliefs can be broken down into sections, separating the description of the background and set-up from the specific sentence that describes or suggests a character’s mental states. Thinking about other minds can thus be pinned to a specific segment within an ongoing story. The right temporo-parietal junction, in particular, shows activity at the point within a single story when a character’s thoughts are mentioned (Mason & Just, 2010; Saxe & Wexler, 2005; Young & Saxe, 2008). A similar manipulation, dividing a 60-second story into 20-second segments, only one of which has mental information, has been used in children (Saxe et al., 2009).

In sum, neuroimaging experiments on understanding other minds produce similar results, across a wide range of participants, methods, and materials. Similar experiments have been conducted in Britain, the USA, Japan, Germany, China, the Netherlands, and Italy (e.g. Anna Leshinskaya, personal communication; Moriguchi, Ohnishi, Lane, Maeda, Mori, Nemoto, et al., 2006; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Schnell et al., 2010; Markus van Ackeren, personal communication). The same brain regions are found in participants ranging from 5 years old (Decety et al., 2008; Gweon et al., 2012; Saxe et al., 2009) to at least 65 years old (e.g. Bedny, Pascual-Leone, & Saxe, 2009; Fletcher et al., 1995). These regions are recruited in adults with high functioning autism (Dufour, Redcay, Young, Mavros, Moran, Triantafyllou et al., 2012), and adults who have been completely blind since birth (Bedny et al., 2009); ongoing work in our laboratory suggests they are found in congenitally deaf adults as well. As described above, the same set of regions responds whether the stimuli are presented in text or with pictures, visually or aurally. Participants can be thinking about the thoughts and feelings of a real person or a fictional

character.² The stimuli can include complex narratives, or just a single thought; participants can be instructed to consider others' thoughts and feelings, or be led to do so spontaneously.

The range of tasks, stimuli, and populations make it all the more striking that these experiments converge on the same conclusions. A consistent group of brain regions shows increased metabolic activity across all of these experiments in the "mental" or "theory of mind" condition, namely regions in bilateral temporo-parietal junction, medial parietal/precuneus, medial prefrontal cortex, and anterior superior temporal sulcus. Though this generalization is striking on its own, a key question is: why? Which cognitive process, invoked by all of these diverse tasks, is specifically necessary and sufficient to elicit activity in these brain regions?

Specificity

During the initial discovery of the ToM brain regions, the first experiments (e.g. Fletcher et al., 1995; Gallagher et al., 2000) compared two conditions that differed on multiple dimensions. Compared with the control stories, the stories about false beliefs also included more individual characters, more specific human actions, more implied human emotions, more invisible causal mechanisms, more social roles, more unexpected events, more demand to consider false representations of the world, different syntax, and so on. In fact, an inherent risk of such complex stimuli is that there may be hidden dimensions along which the stimuli grouped into separate conditions differ and that it is these differences, rather than the intended manipulation, that lead to differential brain activity.

Given all these dimensions, how can we infer which are the necessary and sufficient features that led to activity in each region during a given task? One approach is to try to design an experiment that contrasts minimal pairs: stimuli and tasks that differ only in one key dimension, but are exactly identical on all other dimensions. Taking this approach, Saxe, Schulz, and Jiang (2006b) were able to match both the stimulus and the participant's response, using task instructions to change just how the participants *construed* the stimulus. The stimulus was a stick-figure animation of a girl. Modeled after a false-belief transfer (change of location) task, a bar of chocolate moved from one box to another, while the girl either faced toward the transfer or away. In the first half of the experiment, rather than introducing the task as a false-belief task, participants were trained to treat the stick-figure as a physical cue to the final location of the chocolate bar using three rules, including the critical Rule 1: "Facing = last; Away = first. If the girl is facing the boxes at the end of the trial, press the button for the last box. If the girl is looking away from the boxes, press the button for the first box." Participants were accurate in the task, but found it difficult and unnatural. In the second half of the experiment, participants were then told that for Rule 1, another strategy was possible: namely to view the stick-figure as a person and to consider that person's thoughts. Rule 1 was equivalent to judging, based on what the character had seen, where she *thought* the chocolate was. Both ways of solving "Rule 1" generate the same behavioral responses, but only in the second half of the experiment were participants construing Rule 1 as referring to a character's thoughts. We found that, though the stimuli and the responses were identical across tasks, right TPJ activity was significantly higher in the second half, when participants were using ToM to solve the puzzle, rather than the simple association rule.

² Interestingly, participants can also be assigning hypothetical thoughts and preferences to themselves (e.g. Lombardo et al., 2010; Vogeley et al., 2001). Note, however, that not all metacognition elicits ToM activity. The link between attributing hypothetical thoughts to the self, vs. other kinds of metacognition, is not completely clear (see Saxe & Offen, 2009).

Another approach to dealing with the many dimensions of ToM stimuli is to systematically vary or match each of these dimensions in a long series of experiments. Although each experiment has many differences between the mental and control conditions, across the whole set of experiments, most other kinds of differences are eliminated, leaving only one systematic factor: thinking about thoughts.

For example, Gallagher et al. (2000) showed that none of the low-level features of the original verbal stimuli (e.g. number of nouns, number of straight edges, retinal position) are **necessary** to elicit activity in these brain regions, because they found the same patterns of activity in response to verbal false belief stories and non-verbal false belief cartoons. Within verbal stories, it is not necessary to explicitly state a character's thoughts or beliefs: there is activity in these regions both when people read about a character's thoughts and when they infer those thoughts from the character's actions (Mason & Just, 2010; Young & Saxe, 2009a). Nor is it necessary that the beliefs in question be false: true beliefs, false beliefs, and beliefs whose veracity is unknown are all sufficient to elicit robust activity in these brain regions (Döhnelt, Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., & Sommer, 2012; Jenkins et al., 2010; Young, Nichols, & Saxe, 2010c).

Other experiments showed that the presence of a human character in the stimuli is not sufficient: stories that describe a character's physical appearance, or even their internal (but not mental) experiences, like hunger or queasiness or physical pain, elicit much less response than stories about the character's beliefs, desires, and emotions (Bedny et al., 2009; Bruneau et al., 2011; Lombardo et al., 2010; Saxe & Powell 2006). It is also not sufficient for a story to describe invisible causal mechanisms (like melting and rusting, Saxe & Kanwisher, 2003), unexpected events (like a ball of dough that rises to be as big as a house, Young, Dodel-Ferer & Saxe, 2010b; Gweon et al., 2012), or people's stable social roles (including kinship and professional relationships, Saxe & Wexler, 2005; Gweon et al., 2012), if the story does not also invoke thinking about a person's thoughts.

One particularly important dimension to test was whether considering any representation of the world, mental or otherwise, would be sufficient to elicit activity in these brain regions. Understanding other minds often requires the ability to suspend one's own beliefs and knowledge, and consider the world as it would seem from another perspective. These cognitive processes have been called meta-representation (the ability to conceive of distinct representations of the world, Aichhorn, Perner, Weiss, Kronbichler, Staffen, & Ladurner, 2009; Perner, 1991; Perner et al., 2006), and decoupling (the ability to suspend one's own knowledge in order to respond from a different perspective, Leslie & Frith, 1990; Liu, Sabbagh, & Gehring, 2004). Since meta-representation and decoupling are such essential ingredients of understanding other minds, and especially understanding false beliefs, many scientists initially hypothesized that brain regions recruited by false belief tasks most likely performed one of these two functions. To test this hypothesis, we need stimuli or tasks that require meta-representation and decoupling, but are not about understanding other minds. Currently, the best such example are stories about "false signs" and "false maps" (Zaitchik, 1990). Like beliefs, signs and maps represent (and sometimes misrepresent) reality. Thinking about the world as depicted in a map requires the capacity for meta-representation; when the map is wrong, reasoning about the world as it seems in the map requires decoupling from one's own knowledge of reality. Nevertheless, stories about false signs and maps elicit much less activity in these brain regions (especially right TPJ) than stories about beliefs (Aichhorn et al., 2009; Perner et al., 2006; Saxe & Kanwisher, 2003).

In sum, tasks and stimuli that require, or robustly suggest, thinking about thoughts lead to activity in these regions. Thinking about thoughts can be manipulated by changing participants' instructions for the same stimuli, or by changing the stimuli with the same instructions. Very similar stimuli and tasks, however, which focus on physical objects, physical representations, or externally observable properties, do not lead to activity in these regions.

Links to behavior

The review in the previous sub-section shows that tasks and stimuli that evoke thinking about thoughts also elicit metabolic activity in the ToM brain regions. However, there is even stronger evidence for a link between activity in these regions and understanding other minds: across trials, across individuals, and across development, performance on behavioral tests of ToM is related to brain activity in these same regions.

Most adults pass standard laboratory ToM tasks 100% of the time, leaving little room for inter-individual variability in accuracy. However, by using tasks with no simple right answer, it is possible to reveal individual differences in mental state attribution. Imagine, for example, learning about a girl Grace, who was on a tour of a chemical plant. While making coffee, Grace found a jar of white powder, labeled “sugar,” next to the coffee machine. She put the white powder, which was actually a dangerous toxic poison, in someone else’s coffee. They drank the poison and got sick. Is Grace morally blameworthy? These scenarios require weighing what Grace intended (her mental state) against what she did (the outcome). Participants disagree in their judgments; some people think she is completely innocent (because she believed the powder was sugar), whereas others assign some moral blame (because she hurt someone). This difference is correlated with neural activity during moral judgments across individuals; the more activity there was in a participant’s right TPJ, in particular, the more the participant forgave Grace for her accidental harms (Young & Saxe, 2009b).

An alternative strategy is to measure the quantity and quality of people’s spontaneous mentalistic attributions to ambiguous stimuli. For example, when viewing the simple animations of a small and large moving triangle (Castelli et al., 2000), people generate very rich mentalistic interpretations from the simple movements depicted in these stimuli (e.g. “the child is pretending to do nothing, to fool the parent”). People differ in the amount, and appropriateness, of the thoughts and feelings that they infer from the animations. People who have more activity in ToM brain regions, while watching the animations give more appropriate descriptions of the triangles’ thoughts and feelings after the scan (Moriguchi et al., 2006). Relatedly, Wagner, Kelley, & Heatherton (2011) showed participants still photographs of natural scenes, approximately a quarter of which contained multiple people in a social interaction. Participants performed an orthogonal categorization task (“animal, vegetable, mineral?”). Individuals who scored high on a separate questionnaire, measuring tendencies to think about others’ thoughts and feelings (the “empathy quotient”; Baron-Cohen & Wheelwright 2004; Lawrence, Shaw, Baker, Baron-Cohen, & David, 2004) also showed higher activity in mPFC in response to photographs of interacting people.

Differences in ToM are easier to find in young children, who are still learning how to understand other minds. Interestingly, two recent studies from our laboratories suggest that, overall, getting older, and getting better at understanding other minds, is associated not with more activity in ToM brain regions but with more selective activity. Children from 5 to 12 years old all have adult-like neural activity when listening to stories about characters’ thoughts and feelings. What is different is that ToM regions in younger children show similarly high activity when listening to any information about characters in the story, including the characters’ physical appearance or social relationships (Saxe et al., 2009; Gweon et al., 2012), whereas in older children and adults, the ToM brain regions are recruited only when listening to information about thoughts and feelings (Saxe & Powell, 2006; Saxe et al., 2009). This developmental difference in the selectivity of the ToM brain regions is correlated with age, but also with performance outside the scanner on difficult ToM tasks (Gweon et al., 2012). Moreover, the correlation between neural “selectivity” and behavioral task performance remains significant in the right TPJ, even after accounting for age.

One limitation of all the foregoing studies is that they are necessarily correlational. The strongest evidence that some brain regions are involved in a cognitive task is to show that disrupting those regions leads to biases or disruption in task performance. Transcranial magnetic stimulation (TMS) offers a tool for temporarily disrupting a targeted brain region. We (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010a) compared people's moral judgments following TMS to either right TPJ or a control brain region 5 cm away. TMS to right TPJ, but not the control region, produced moral judgments temporarily biased away from considerations of mental state information. Innocent accidents appeared more blameworthy, while failed attempts appeared less blameworthy, as though it mattered less what the agent believed she was doing, and more what she actually did. People didn't lose the ability to make moral judgments altogether; they still judged that it is completely morally wrong to intentionally kill, and not wrong at all to simply serve someone soup. Disrupting the right TPJ thus appears to leave moral judgment overall intact, but impairs people's ability to integrate considerations of the character's thoughts into their moral judgments. Converging evidence comes from another TMS study: TMS to right TPJ made adults slower to recognize a false belief in a simple (non-moral) false belief task (Costa, Torriero, & Oliveri, 2008).

Another way to study the necessary contributions of a brain region is to work with people who have suffered permanent focal (i.e. local) damage to that region, typically due to a stroke. Samson, Apperly, and colleagues (Apperly & Butterfill 2009; Apperly, Samson, & Humphreys, 2005; Samson, Apperly, & Humphreys, 2007) have conducted a series of elegant studies using this approach. Initially, these authors tested a large group of people, with damage to many different brain regions, on a set of carefully controlled tasks. They then identified individuals with a specific pattern of performance: individuals who passed all the control tasks (e.g. measuring memory, cognitive control, etc.), but still failed to predict a character's actions based on their false beliefs. Next, the scientists used a lesion-overlap analysis to ask which brain region was damaged in all, and only, the patients with this diagnostic profile of performance. The answer was the left TPJ, one of the same brain regions identified by fMRI.³

Summary

The literature from the last 10 years thus suggests a generalization—there are cortical regions in the human brain where activity is associated with understanding other minds in three ways:

1. Metabolic measures of activity reliably increase when the participant is thinking about thoughts, across a wide range of stimuli and tasks, but not in response to a variety of similar control tasks and stimuli.
2. Activity is correlated with behavioral measures of thinking about thoughts.
3. Disrupting activity leads to deficits in thinking about thoughts.

So far, these claims are relatively uncontroversial. As we noted above, there is a broad consensus in social cognitive neuroscience. However, much controversy remains about the proper interpretation of these data.

Exactly what is the nature of these regions, their functions, and their contribution to thinking about thoughts? Here's a strong hypothesis: one or more of these regions has the specific

³ It is worth noting that the effects of lesions to the right TPJ, one of the regions argued to be most selective for ToM, haven't yet been effectively tested. The candidate participants all had extensive and diffuse damage to the right hemisphere, and failed the control tasks, making it impossible to test ToM deficits specifically.

cognitive function of representing people's mental states and experiences—that is, of thinking about thoughts. Whenever we are thinking about thoughts, there are neurons in these regions firing. These neurons are gathered in spatial proximity (i.e. into a “region”) because they have related computational properties that are distinct from the computation properties of neurons in the surrounding cortex.⁴ The pattern of firing, in space and time, of these neurons encodes aspects of someone's thoughts. As an analogy, consider the way MT neurons encode speed and direction of motion, and face area neurons encode aspects of facial features that are relevant to face identity (Freiwald, Tsao, & Livingstone, 2009; Georgopoulos, Schwartz, & Kettner, 1986). In the proposed hypothesis, the neurons in the ToM brain regions encode aspects and dimensions of inferred thoughts. Scrambling the pattern of activity in these neurons would therefore lead to an inability to discriminate one inferred mental state from another, for example, making all minds appear homogenous: people might all seem to have the same desires and preferences, and the same knowledge and beliefs. More serious damage to these regions might make it impossible to think about other minds at all, without similarly impairing the rest of cognition.

There certainly is not enough evidence to prove that this strong hypothesis is right; a more immediate question is whether it is obviously wrong. There are at least two classes of potential objections: theoretical arguments, based on general principles of how the brain works, and empirical arguments, based on the results of other experiments in cognitive neuroscience. In the next section, we describe some of these objections, and some of our responses to them.

A strong hypothesis

Objections from theoretical considerations

Many authors have expressed discomfort with the project of trying to link specific cognitive functions with delineated brain regions. For example, a decade after their 2nd edition UoM chapter, Chris and Uta Frith wrote: “We passionately believe that social cognitive neuroscience needs to break away from a restrictive phrenology that links circumscribed brain regions to underspecified social processes” (Frith & Frith, 2012). Others have echoed this accusation of phrenology; for example, Bob Knight criticizes the “phrenological notion that a given innate mental faculty is based solely in just one part of the brain” (Knight, 2007), and William Uttal recently argued that “any studies using brain images that report single areas of activation exclusively associated with any particular cognitive process should a priori be considered to be artifacts of the arbitrary thresholds set by investigators and seriously questioned” (Uttal, 2011). Most such theoretical objections include variations on three themes: social cognitive neuroscientists are accused of (incorrectly) viewing regions as (1) functioning in isolation, (2) internally functionally homogenous, and (3) spatially bounded and distinct. Here, we address each of these concerns in turn.

First, does claiming that a region has a specific function (e.g. in thinking about thoughts) entail suggesting that this region functions in isolation? To put it more extremely, are we claiming that, for example, the right temporo-parietal junction (RTPJ) could pass a false belief task on its own? Obviously not. Performing any cognitive task necessarily depends on many different cognitive and computational processes, and therefore brain regions. No interesting behavioral task can be accomplished by a single region. The tasks of the mind and brain—recognizing a friend, understanding a sentence, deciding what to eat for dinner—must all be accomplished by a long sequence

⁴ These distinct properties may derive entirely from patterns of connectivity, not from the structure of the neurons themselves.

of processing steps, passing information between many different regions or computations, from sensory processing all the way to motor action. The function of a neuron or a brain region should never be identified with completing a cognitive task. Thus, for example, “passing a false belief task” is not even a candidate function of a brain region. Any time an individual passes a false belief task, many brain regions—involved in perceiving the stimuli, manipulating ideas in working memory, making a decision, and producing a response—will all be required (Bloom & German, 2000).

More generally, we expect that the functions of regions (or neural population, regardless of spatial organization) will be to receive a class of inputs, and transform them into output, which make different information relatively explicit. Therefore, the specific questions about any neural population should include: what input does it receive, what output does it produce, and what information is made explicit in that transformation? Of course, the answers to these questions will require us to understand the position of this neural population within a larger network, especially when characterizing a region’s input and output. In the case of the ToM regions, a related question concerns the relationships between the different regions within the network. At least five cortical regions are commonly recruited during many different social cognitive tasks: how is information passed between, and transformed by, each of these spatially distinct regions?

Thus, studying the function of a brain region means studying in isolation one component of a system that could never function in isolation. This description may sound ominous, but scientific progress frequently requires us to break complex systems into component parts. While the pieces could not function in isolation, understanding their isolated contributions is necessary to understanding the function of the integrated system. For any given neural population, it is reasonable to ask: what classes of stimuli and tasks predictably and systematically elicit increased activity in the population as a whole? Which dimensions of stimuli lead to activity in distinct subpopulations of neurons? Both traditional and new fMRI methods help answer these questions, albeit somewhat indirectly (more on this, in “Where next?”).

The second objection is that studying brain regions leads to the false assumption that groups of spatially adjacent neurons are functionally homogenous. The regions we study in fMRI are orders of magnitude larger than what we believe are the true computational units of brain processing, the neurons. Changes in blood oxygenation measured by fMRI inevitably reflect averages over the activity of many thousands of individual neurons. Why is it useful to study oxygen flow to chunks of cortex approximately 1–5 cm² in size, which are so much bigger than neurons, but so much smaller than the networks required to complete a task?

Our suggestion is that there is no reason, *a priori*. It just happens, as a matter of empirical fact, that many interesting computational properties of the brain can be detected by studying the organization of neural responses on this scale. Aggregating the responses of neighboring neurons often produces informative population averages. Results obtained via fMRI reflect the same distinctions found in directly observable population codes (e.g. Kamitani and Tong, 2005; Kriegeskorte & Bandettini, 2007). This empirical fact may have a theoretical explanation. Neurons with similar or related functions may be spatially clustered to increase the computational efficiency of frequent comparisons (e.g. lateral inhibition). Blood-oxygen delivery to the cortex may follow the contours of neural computations, to increase the hemodynamic efficiency of simultaneously getting oxygen to all of the neurons that need it (Kanwisher, 2010). However, these arguments are not necessary premises; for fMRI to be useful, we only need the empirically observable fact that useful and reliable generalizations can be made for hemodynamic activity in patches of cortex at the spatial scale of millimeters.

Of course, though, neurons within a region or an fMRI voxel are never functionally homogenous. Consider the analogy of primary visual cortex—neurons in V1 have visual receptive fields,

meaning that activity can be induced by a pattern of bars of light falling on a specific region of the retina. However, neurons in V1 differ from one another in where on the retina one should place the light (retinotopy), how large the pattern should be (size and spatial frequency preferences), and the orientation to which the bars should be rotated (orientation selectivity), to elicit a maximal response. There is also a separate (but systematically interleaved) population of neurons for which the response depends on color, but not orientation or size. Furthermore, some neurons primarily send information to subsequent regions of visual cortex (e.g. excitatory neurons) whereas other neurons primarily modulate the response of neighboring neurons in V1 (e.g. inhibitory interneurons). As far as we know, the metabolic activity measured by fMRI reflects a combination of activity in all of these different populations. Consequently, we should never assume that the amount of “activity” we measure in a region with fMRI represents (or would correlate very well with) the rate of firing of any individual neuron inside that region. Similarly, we cannot assume that if two different stimuli or tasks elicit similar magnitudes of activity in a region, then they are eliciting responses in the same, or even shared, neural populations. Completely non-overlapping subpopulations of neurons could produce the same magnitude of fMRI activity within a region. Any interpretation of fMRI data must be sensitive to this possibility. In fact, studying the organization of functional subpopulations within a region (e.g. which dimensions of stimuli are represented by distinct subpopulations within a region) may be one of the most powerful ways that fMRI will contribute to the neuroscience of ToM. We describe these methods in greater detail in “Where next?”

The third potential objection is that studying regions with fMRI leads researchers to imagine boundaries between discrete regions, when the truth is a continuous distribution of neural responses over the cortical sheet. The data we described in the first section shows that cortex is not functionally homogenous with regard to theory of mind, and regional distinctions are not all “artifacts of arbitrary thresholds.” Still, there is a legitimate reason why cognitive neuroscientists may be reluctant to call any reliable functional regularity discovered by fMRI a “region.” These “regions” may turn out to be just one piece of a larger continuous functional map over cortex, not computationally distinct areas of their own (Kriegeskorte, Goebel, & Bandettini, 2006).

Cortical responses at scales measurable by fMRI are organized along multiple orthogonal spatial principles. One is the division of cortex into cytoarchitectonic “areas,” like primary visual cortex (V1) and primary auditory cortex (A1). Orthogonal to the division of cortex into areas are topographic principles. Most visual regions, for example, are organized by retinotopy: moving across the cortical sheet, the region of the visual field eliciting a maximal response varies smoothly, covering the whole visual field from fovea to periphery, top to bottom, and left to right. Likewise, multiple distinct motor areas are organized by somatotopy, and auditory areas by tonotopy.

These orthogonal principles of cortical organization create a challenge for cognitive neuroscientists, because in charting new territory, away from well-understood sensory and motor systems, we may claim to discover new functional regions associated with higher-order cognitive processes, which are really just one end of a larger map (cf. Konkle & Oliva, 2012). To make the puzzle concrete, imagine looking at functional responses to visual stimuli across occipital cortex for the first time, without the benefit of the history of visual neuroscience. One tempting way to divide the occipital cortex into functional “regions” might be by retinotopy—one group of patches that responds to foveal stimuli, and a different group of patches that responds to peripheral stimuli. This foveal vs. peripheral difference is highly robust, replicable within and across subjects, within and across tasks, and correlated with behavior (i.e. visual performance in corresponding regions of the visual field). Nevertheless, other considerations, such as cytoarchitecture, connectivity, and processing time, suggest that this is the wrong division for capturing functional and computational regularities. Identifying a robust functional regularity that divides one patch of cortex from

another is not the same thing as identifying a true cortical area—a region that is computationally distinct from its neighbors, with distinct cytoarchitecture, connectivity, and topography (Friston, Holmes, Worsley, Poline, Frith, & Frackowiak, 1995; Kanwisher, 2010; Kriegeskorte et al., 2006; Worsley, Evans, Marrett, & Neelin, 1992). Instead of studying the “peripheral patches,” neuroscientists divide the cortex into V1, V2, V3, MT, etc., each of which is then internally organized (in part) by retinotopy.

Imagine you are a social cognitive neuroscientist, looking at a new bit of cortex, and you see a new functional regularity—a patch of cortex that shows a high response when the individual is thinking about stories, cartoons, or movies depicting the contents of another mind. Which kind of functional regularity is this? On the one hand, these data could signal the discovery of a true computational area, like V1. On the other hand, the observed functional regularity might be more like “peripheral patches,” one part of a stimulus space or dimension that is mapped across cortex, but crosscuts multiple computational areas.

We believe that current fMRI data cannot resolve this puzzle directly. One approach may therefore be to suspend judgment until other sources of evidence are available. If patches of cortex involved in understanding other minds are true cortical areas, it will be possible to distinguish them from their anatomical neighbors by cytoarchitecture, connectivity, and/or topography. Non-invasive functional imaging tools do not yet have high enough resolution to reveal cytoarchitecture *in vivo*. To study the links between function and cytoarchitecture with current technology, we would need to collect functional data to identify the regions *in vivo*, and then analyze the neuroanatomy of the same individual post mortem.

A weaker, but more accessible, source of evidence is patterns of connectivity. In some cases, cortical areas can be differentiated by their profiles of connectivity. It has become increasingly possible to measure the pattern of connections between regions using neuroimaging. Diffusion imaging (DTI), which looks at the predominant direction of water diffusion, allows us to visualize the dominant pathways of axons connecting brain regions. Using diffusion, we can ask whether a patch of cortex involved in understanding other minds shows a different pattern of connectivity than its neighbors to the rest of the brain. Initial evidence suggests it does, at least in the case of the right TPJ. Mars, Sallet, Schüffelgen, Jbabdi, Toni, & Rushworth (2012) found that the broad area of right temporo-parietal cortex (BA 39/40) can be sub-divided into three clusters, based on DTI connectivity alone, and one of these clusters is functionally correlated (during a resting baseline) with the other ToM regions, including medial prefrontal and medial parietal cortex. Although Mars and colleagues did not directly test whether the region defined by connectivity and the region defined by function (i.e. active in ToM tasks) share the same boundaries, the results are suggestive.

Currently, however, neither cytoarchitecture nor connectivity analyses give a definitive evidence that patches of cortex recruited by mental state reasoning tasks are true cortical areas. An alternative approach might therefore be to consider the alternative hypothesis directly—that these regions are one part of a larger, continuous map. If we compare the cortical patches we are studying to their anatomical neighbors, is there a plausible higher-level “stimulus space,” which could unite these responses into one map?

Again, the right TPJ is an interesting example. The right TPJ region that is activated by ToM tasks (as described in “Theory of mind and the brain”) has two very close anatomical neighbors. One neighbor (up toward parietal cortex in the right inferior intraparietal sulcus (IPS), but confusingly sometimes also called RTPJ) is recruited by unexpected events that demand attention. These events may be unexpected because they are rare, or because a generally reliable cue was misleading on this occasion. Redirecting attention toward an unexpected event leads to metabolic activity in this region (Corbetta & Shulman, 2002; Mitchell, 2008; Serences, Shomstein, Leber,

Golay, Egeth, & Yantis, 2005). Damage to this region makes it hard for objects and events in the contralateral visual field to attract attention, producing left hemifield neglect (Corbetta, Patel, & Shulman, 2008). Some authors have noted that false belief tasks typically involves an unexpected event (Corbetta et al., 2008; Decety & Lamm, 2007; Mitchell, 2008): for example, false belief tasks often hinge on an object unexpectedly changing location while the protagonist is out of room, and require the participant to shift attention between both locations. In fact, the region that is recruited during exogenous attention tasks is so close to the region recruited during ToM tasks that some concluded that these are actually just two different ways to identify the same region (Corbetta, Kincade, Ollinger, McAvoy & Shulman, 2000; Mitchell, 2008). More recently, however, both meta-analyses and high-resolution scanning within individual subjects suggest that there are actually two distinct cortical regions (or “patches”), and the region recruited by attention tasks is approximately 10 mm superior to the region recruited by ToM tasks (Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009; Decety & Lamm, 2007). Also, the region that is recruited during false belief tasks is not recruited by unexpected transfers of location in stories about false maps and physical representations, as described above (e.g. Young et al., 2010b). Nevertheless, it remains an interesting possibility that the exogenous attention response and the ToM response are part of a larger continuous map across cortex, a topography of different kinds of unexpected attentional shifts. The more superior end of this map could direct attention toward unexpected positions in space and time, while the more inferior end of the map could direct attention toward unexpected people, actions, or inferred thoughts.

A second topographical “stimulus space” of which the RTPJ could be a part runs anteriorly, through the right superior temporal sulcus (STS), and down toward the temporal pole. As with the RTPJ and the right IPS, the RTPJ and the posterior STS were initially conflated, but have subsequently been shown to be spatially distinct (Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007). Multiple parts of the STS are recruited when participants observe other people’s actions in photographs, film clips, point light walkers or animations (Pelphrey, Mitchell, McKeown, Goldstein, Allison, & McCarthy, 2003; Pelphrey, Morris, Michelich, Allison, & McCarthy, 2005; Hein & Knight, 2008). The STS is large, and Kevin Pelphrey and colleagues (Pelphrey et al., 2005; Pelphrey & Morris 2006) propose that it contains a pseudo-somatotopic map of observed actions: others’ mouth motions represented most anteriorly, followed by hand and body movements, followed by head and eye movements represented most posteriorly. An intriguing possibility is that the temporo-parietal junction, which is at the most posterior end of the superior temporal sulcus, is part of this same map. Whereas hand and body movements convey information about what a person is doing and intending, head and eye movements convey information about what a person is looking at and seeing. Thus, the STS may contain a map of others actions that move from externally observable body movements (anterior end) toward invisible mental states (posterior end), culminating in the RTPJ, which responds to thinking about what a person is thinking.

In principle, either of these mapping hypotheses, or both, could be true. Evidence that a ToM region, e.g. the RTPJ, is part of a larger cortical map might come in the form of a response that moves continuously across contiguous patches of cortex, modulated by continuous changes in a stimulus dimension (Konkle & Oliva, 2012). For example, if the RTPJ is one end of an attention map, we might expect to see that surprising facts about physical entities elicit activity relatively far from the TPJ, but that as the unexpected stimulus becomes more social and interpersonal, activity moves continuously across the map, ending at the RTPJ. Similarly, if the RTPJ is the “abstract” or “head” end of a map of observable social actions, we’d expect to see continuous changes in the location of activity, as depictions of human actions become either more abstract or physically higher on the body. Finally, both could be true. The RTPJ could exist at its precise location because

that is where a region that deals with unexpected information converges with a region that deals with human actions. We would be enthusiastic to see someone test these hypotheses further.

For now, however, we suggest that it does not matter for any of our empirical purposes whether the RTPJ (or any other patch of cortex recruited in ToM tasks) is a true cortical area, or whether it is one end of a larger map, as defined by cytoarchitecture, connectivity, and topography. In either case, there is a robust functional regularity replicable within and across subjects, within and across tasks, and correlated with behavior: a bounded, adjacent patch of cortex where activity is high during a range of ToM tasks. Without committing to whether they are true computational “areas,” or just the equivalent of “Peripheral Patches,” we believe that it is fruitful to continue to study these coherent patches as “regions,” in order to discover the computations and representations that underlie thinking about other minds.

Given everything else we know about the brain, it is not surprising that systematic response profiles are linked to regions of cortex much bigger than a neuron and much smaller than a network. The challenge now is to integrate the huge, and growing, list of empirical discoveries and to construct hypotheses about the computations and representations in the neural populations we are studying. These empirical results form the basis of a different set of potential objections to the hypothesis that there is a strong specific link between cortical regions, like the TPJ, mPFC, and PC, and thinking about thoughts.

Empirical objections

An empirical objection to our argument in “Theory of mind and the brain” might begin by pointing out that our review of the literature was selective. In addition to the dozens of articles we cited, there are dozens of others that claim so-called “ToM regions” are active in tasks that do not involve thinking about thoughts, or are not active in tasks that do involve thinking about thoughts. How can we integrate these other data into a coherent hypothesis?

First, what about claims that “ToM regions” are active in tasks that do not involve thinking about thoughts? The RTPJ is again a useful example. As we mentioned above, some authors initially believed that the same region of RTPJ involved in false belief tasks was also recruited during any exogenous shift of attention and/or biological motion perception. Other literature suggests that the RTPJ is involved in maintaining a representation of one’s own body, by integrating multi-sensory information and locating the body in space (Blanke et al., 2005; Blanke & Arzy, 2005; Tsakiris et al., 2008). Experiments that ask people to mentally rotate their own body or imagine their body in different parts of space, as well as those that induce changes in bodily self-perception (with e.g. rubber hands) find activation in this region (Arzy, Thut, Mohr, Michel, & Blanke, 2006; Blanke & Arzy, 2005). TMS and intracranial stimulation to this region has been shown to lead to out-of-body experiences, confusion of the body vs. the environment, and illusory changes in the orientation of body parts (Blanke et al., 2002, 2005; Tsakiris et al., 2008).

How do we interpret these results, which seem to contradict our theory? In general, we could consider five possibilities. The first is that one group of scientists has made an error, leaving an unnoticed confound in their experimental paradigm. Could the body representation tasks also involve thinking about thoughts? Could the false belief task accidentally induce updating maintaining a representation of one’s own body? These are empirical questions, but we consider both possibilities highly unlikely. The second option is that there is some deep common computation, served by the same neural population that is required by these different classes of tasks. One option here would be “decoupling” (Gallagher & Frith, 2003; Leslie & Frith 1990; Liu, Sabbagh, & Gehring, 2004), maintaining distinct representational reference frames, e.g. for one’s own and other minds, for current vs. imaginary body positions, and so on. The third option is that the same

neurons can assume distinct and even unrelated functional roles, depending on the context and the pattern of activity in other neural populations (e.g. Miller & Cohen, 2001). On this view, thinking about thoughts and maintaining a body representation are cognitively unrelated, in spite of being implemented by the same neurons. The fourth option is that distinct neuronal populations are involved in thinking about thoughts and maintaining a body representation, but these neurons are interleaved within the RTPJ, just as color- and orientation-sensitive neural populations are interleaved in V1.

Finally, the fifth option is that distinct neuronal populations are involved in thinking about thoughts vs. maintaining one's body representation, exogenous attention shifts, and biological motion perception. These neural populations are not interleaved: they are contained in distinct regions that are merely nearby on the cortical sheet. We find this last option most plausible. Standard fMRI methods, which involve extensive spatial blurring at three stages (acquisition, preprocessing, and group averaging; Fedorenko & Kanwisher, 2009; Logothetis, 2008), are strongly biased to conflate neighboring regions that are truly distinct. Even so, the existing evidence suggests that the region involved in body representations is lateral (MNI x-coordinates typically around 64mm) to the region involved in thinking about thoughts (x-coordinates around 52 mm). As described above, this was also true of the regions involved in biological motion perception (Gobbini et al., 2007) and exogenous attention (Decety & Lamm, 2007; Scholz et al., 2009); these are almost completely non-overlapping in individual subjects.

A different kind of challenge arises from examples of tasks that apparently do involve thinking about thoughts, but do not elicit activity in TPJ, mPFC, or PC. Two interesting examples are visual perspective-taking tasks, and recognition of facial expressions of emotions from photographs. In visual perspective taking tasks, the participant sees an image of a character in a 3D space, and is asked to imagine the view of the room from the viewpoint of the character. For example, participants may be asked to report how many dots the character can see (the third person perspective), versus how many dots the participants themselves can see (including those that are out of the character's view, the first person perspective). This perspective-taking task clearly involves thinking about the character's visual access, which could be construed as a mental state. Nevertheless, this task typically does not elicit activity in the same regions as thinking about thoughts (Aichhorn et al., 2006; Vogeley May, Ritzl, Falkai, Zilles, & Fink, 2004). Similarly, participants in hundreds of fMRI experiments have viewed photographs of human faces expressing various basic emotional expressions (e.g. sad, afraid, angry, surprised, happy, neutral). Although these images do depict evidence of another person's emotional experience, they also typically do not elicit activity in the same regions as thinking about thoughts (Costafreda et al., 2008; Lamm, Batson, & Decety, 2007; Vuilleumier et al., 2001).

What should we conclude from these examples? One option is to make a forward inference, using the evidence from these tasks to change our hypothesis about the brain regions' functions. Here, the forward inference might be that the ToM brain regions are responsible for only a subset of mental state processing. We could conclude that different brain regions are involved in thinking about different classes of internal experiences: bodily states, emotional states, perceptual states, or epistemic states (e.g. thinking, knowing, doubting, etc.). The so-called ToM brain regions might be specifically involved in representing epistemic states, while regions of insula represent others' emotions and regions of parietal cortex represent others' perceptual states.

Another option is to make a reverse inference: using the pattern of neural activity to change our analysis of the cognitive processes required by the task. Reverse inferences are risky because they require a lot of confidence in the functional specificity of the brain region(s) involved (Poldrack, 2006a). However, we think ToM brain regions are good candidates to support reverse inference,

given the converging evidence across the many experiments described above. In this case, a reverse inference might be that these visual perspective taking and emotional face tasks do not actually elicit thinking about thoughts, and instead are solved by alternative computational strategies. For example, rather than truly considering someone's perceptual experiences, line-of-sight tasks may be solved using mental rotation and geometric calculation. Emotional facial expressions may be recognized (akin to object recognition) without always requiring a representation of the person's internal state.

In these particular cases, we are open to either the forward or reverse inference; only further experimentation will tell which is the better generalization. In general, we believe that in this domain, inferences can be made in both directions, forward and reverse. The absence of activation in perspective taking and emotion recognition tasks provides an important constraint on the possible functions of ToM brain regions (the forward inference). At the same time, the fact that visual perspective-taking and emotion recognition rely on different brain regions from reading stories about thoughts provides evidence that these tasks depend on different cognitive processes (the reverse inference). It's the give and take of these two kinds of inferences, as evidence accumulates, that allows us to build a coherent understanding of both cognitive and neural function.

Summary

Taken together, these first two sections illustrate the main contributions of the first decade of neuroimaging the understanding of other minds. We have discovered a robust, replicable functional regularity in the human brain: regions that have increased activity when participants think about thoughts. These regions may be true cortical areas or parts of larger topographical maps, but in either case, understanding other minds is a major organizing principle of responses over cortex. The function of these regions is not to complete a task, but to transform some class of input into some output; and the class of input has something to do with thinking about minds, and not bodies or abstract representations. Of course, this description remains unsatisfying and largely underspecified. Nevertheless, we are optimistic that the second decade of this research program will continue to improve our specifications. In particular, we are excited about newly emerging methods for fMRI data analysis that focus on the second aspect of a region's function: the features within its preferred stimulus class that organize differential responses within each of the ToM regions.

Where next?

Differences between theory of mind regions

One step in specifying the computations performed by ToM regions will be understanding the division of labor and information transfer between the different regions. Overall, ToM regions show similar profiles to most of the contrasts we described. However, research in the last 5 years has begun to tease apart the functional profiles of these regions, and the differences are intriguing, though much work still remains to be done to form a coherent view of how they all fit together.

The most striking contrast comes from a task that elicits very robust activity in the medial ToM regions (mPFC and precuneus), but not the lateral ToM regions (TPJ and anterior STS): thinking about personality traits, especially of the self and close others (Whitfield-Gabrieli et al., 2011; Moran et al., 2011a; Saxe, Moran, Scholz, & Gabrieli, 2006a; Krienen, Tu, & Buckner, 2010). In a typical version of the experiment, participants in the scanner see single words describing personality traits (e.g. "lazy," "talkative," "ambitious") and judge whether each one is desirable or undesirable

(the semantic control condition), is true of a famous person (the other control condition), or is true of themselves (the self condition). MPFC and precuneus regions show much higher activity during the self condition; moreover, within the self condition, activity in mPFC is higher for the words that participants say **are** true of themselves (Moran et al., 2011a), and the amount of activity in mPFC for each item presented in the self task (but not in the semantic or other tasks) predicts participants' subsequent memory for those items on a surprise memory test (Mitchell, Macrae, & Banaji, 2006; Jenkins & Mitchell, 2009). These data are compelling to us: the mPFC, but not the TPJ, is involved in reflection about one's own stable traits and attributes (Lombardo et al., 2010). Similarly, elaborating one's own autobiographical memories leads to activity in medial ToM regions, whereas imagining someone else's experiences on similar occasions elicits activity in bilateral TPJ (Rabin, Gilboa, Stuss, Mar, & Rosenbaum, 2010).

In fact, coding information in terms of similarity to the self may be a key computation of mPFC. In one series of studies (Tamir & Mitchell, 2010), participants judged the likely preferences of strangers (e.g. is this person likely to "fear speaking in public" or "enjoy winter sports") about whom they had almost no background information. Under those circumstances, the response of the mPFC (but not TPJ) was predicted by the discrepancy between the attributions made to the target and the participant's own preference for the same items: the more another person was perceived as different from the self, for a specific item, the larger the response in mPFC.

Another distinction, supported by multiple studies, suggests that sub-regions of mPFC are most recruited when thinking about someone's negative emotions or bad intentions, whereas the TPJ makes no distinction based on valence. For example, Bruneau, Pluta, & Saxe (2011) found that only the mPFC showed a higher response to stories about very sad events (e.g. a person proposes marriage and is rejected) compared to neutral or positive events (e.g. the marriage proposal is accepted). In a PET study, Hayashi, Abe, Ueno, Shigemune, Mori, Tashiro, et al. (2010) found that a region in mPFC was recruited when people were considering an actor's dishonesty as a factor in moral judgments; and in an fMRI study, Young & Saxe (2009a) found that a region in ventral mPFC was correlated with moral judgments of attempted harms, which are morally wrong only because of the actor's negative intentions. Converging with these neuroimaging studies, lesion studies suggest that focal damage to ventral mPFC creates disproportionate difficulty in understanding bad intentions, and in integrating those intentions into moral judgments (Koenigs et al., 2007).

In this vein, further work has been done to examine the response profile of mPFC. The "regions" implicated in ToM are very large, especially in the mPFC, and there may be multiple sub-divisions, each with different response profiles. Proposed distinctions along the ventral-dorsal axis of mPFC include: similarity to self (such that self-relevant processes elicit responses more ventrally (e.g. Mitchell et al., 2006; Jacques, Conway, Lowder, & Cabeza, 2011), interpersonal closeness (people who are closer, or more important to the self elicit responses more ventrally, Krienen et al., 2010), or affective content ("hot" affective states elicit responses more ventrally (Ames, Jenkins, Banaji, & Mitchell, 2008; D'Argembeau et al., 2007; Mitchell & Banaji, 2005), while "cool" cognitive states elicit responses more dorsally (Kalbe, Schlegel, Sack, Nowak, Dafotakis, Bangard, et al., 2010; Shamay-Tsoory & Aharon-Peretz, 2007; Shamay-Tsoory, Tomer, Berger, Goldsher, & Aharon-Peretz, 2005). Thus, while there may be a convergent theoretical account of the mPFC, and its responses to other people's negative intentions and emotions, one's own personality traits, and ambiguous inferences about preferences, another possibility is that these response profiles reflect distinct sub-regions within mPFC, each contributing a distinct computation to understanding other minds.

Intriguingly, none of these distinctions have shown to affect the lateral ToM regions. In contrast, one dimension that seems to influence the magnitude of response in TPJ more than mPFC is

whether someone's thought or feeling is unexpected, *given the other information you have about that person*. Saxe & Wexler (2005) introduced characters whose social background was mundane (e.g. New Jersey) or unusual (e.g. a polyamorous cult). Participants then read about that character's thoughts and feelings (e.g. a husband who believed it would be either fun or awful if his wife had an affair). On their own, neither the background nor the content of the belief affected the magnitude of response in the TPJ, but there was a significant interaction: whichever thought was unlikely, given the character's social background, elicited a larger response in the right TPJ. Recently, Cloutier, Gabrieli, O'Young, & Ambady (2011) provided a conceptual replication of this result: participants saw photographs of people labeled as Democratic or Republican, paired with opinions that were either typical of their political affiliation or typical of the opposite affiliation. Opinions that were unexpected given the protagonist's political background (e.g. a Republican wanting liberal Supreme Court judges) elicited a higher response in most of the ToM regions, including bilateral TPJ and mPFC. Finally, a third study suggests that the conflict between background and belief is necessary for increased activation, not just sufficient. In the absence of specific background information about the believer, there is no difference in the response of any ToM region to absurd vs. commonsense beliefs (e.g. "John believes that swimming in the pool is a good way to grow fins/cool off," Young et al., 2010b).

Although they are preliminary, we find these results exciting because they are consistent with the idea that activity in TPJ reflects a process of forming a coherent model of another's mind. We expect other people to be coherent, unified entities, and strive to resolve inconsistencies with that expectation (see Hamilton & Sherman, 1996). Consequently, when a target's behavior violates a previous impression of that person, observers spend more time processing the behavior (Bargh & Thein, 1985; Higgins & Bargh, 1987) and more time searching for the cause of the behavior (Hamilton, 1988). Concomitantly, more activity in TPJ occurs precisely when participants are likely exerting effort to integrate a person's thoughts and feelings into a coherent model of their whole mind; that is, when participants are building a "theory" of a mind (Gopnik & Meltzoff, 1997).

These results suggest that while TPJ activity may be related to the discrepancy between a thought or feeling and other information about the protagonist, mPFC and PC activity may be related to discrepancies between the protagonist's thoughts or feelings and the participant's own thoughts or feelings on the same topic. Thus, whereas TPJ may be involved in integrating a belief or preference into a coherent model of another's mind (e.g. Young et al., 2010c), mPFC and precuneus activity may reflect a different "anchor-and-adjust" strategy, that helps identify other people's thoughts and preferences by starting with one's own preferences, and then adjusting them as necessary (Tamir & Mitchell, 2010). Taken together, these results suggest that medial and lateral ToM regions support distinct computations within ToM. These distinctions help us separate ToM into its real (i.e. neurally-realized) component parts, and formulate hypotheses about each of the more specific functions of individual regions within the group.

Magnitude: "more" theory of mind

Until now, we asked simply whether ToM brain regions do or do not show activity in response to a task or stimulus. However, this is clearly an over-simplification; activation is continuous, not discrete, making it very tempting to ask, "Which stimulus and task dimensions within the domain of thinking about minds modulate the activation in these brain regions?" The magnitude of activity, over tasks or stimuli, could reveal not only what class of stimuli is processed in a region, but also which dimensions of those stimuli and tasks elicit more or less processing.

Interestingly, initial attempts to modulate activation in the ToM network mostly discovered features that do not elicit differential magnitudes of response. For example, most of the ToM regions

show an equally high response to explicit descriptions of beliefs that are true or false (Young & Saxe, 2008), justified or unjustified (Young, Nichols, & Saxe, 2010c), and well-intentioned or bad (i.e. a girl who believes she is putting poison in her friend's coffee, vs. believes that she is putting sugar in the coffee, Young, Cushman, Hauser, & Saxe, 2007). In the TPJ, at least, it also does not matter to *whom* the thought or feeling is attributed: there is an equally strong response to beliefs attributed to similar or dissimilar others (Saxe & Wexler, 2005), or to members of one's own group vs. an enemy group (Bruneau & Saxe, 2010; Bruneau, Dufour & Saxe, 2012).

Part of the reason for this lack of success may be that we do not yet have satisfactory cognitive or computational theories of ToM that allow us to predict when "more" ToM processing will be required, or even exactly what "more" means. Some intuitive possibilities have already proven empirically false. For example, making it harder to infer what a character believes, by making the available evidence more ambiguous, does not lead to more activity in TPJ regions (Jenkins & Mitchell, 2009). In fact, we (Dodell-Feder et al., 2011) found that, while some stories about thoughts systematically elicit more activity than others, in each ToM region, we could not find any feature (e.g. vividness, unexpectedness, length, syntactic complexity) that predicted these differences in activity, with the partial exception of the precuneus, which showed greater activity to stories that involved more people.

Researchers with a background in computer science or game theory often suggest one particular dimension for "more" ToM processing—the depth of embedding of one mental state within another. Thus, many people intuit, reasoning about an embedded belief (e.g. "Carla believes that Ben thinks that she eats too much junk food") should require more ToM processing than reasoning about a simple belief (e.g. "Carla believes that she eats too much junk food.") When we tested this hypothesis directly using verbal stories, we found that no ToM regions showed greater activity for the more embedded beliefs (Koster-Hale & Saxe, 2011). Other brain regions did show more activity—regions involved in language processing and regions involved in difficult memory and cognitive control tasks, like dorsolateral prefrontal cortex (DLPFC). Our interpretation of these results is that embedding beliefs inside other beliefs makes the reasoning problem harder, but does not lead to greater ToM processing, *per se*. This finding converges with results from patient populations and aging adults showing that failure to pass second-order false belief tasks may, in fact, be due to domain-general impairment, rather than diminished ToM processing (Slessor, Phillips, & Bull, 2007; Zaitchik et al., 2006).

This highlights a more general problem with trying to elicit ToM in games. Difficult games often demand ToM reasoning, but similar patterns of behavior can be achieved by logical problem solving. Thus, in games designed to allow for more or less sophisticated ToM reasoning, some papers find ToM regions correlated with increasing "levels of embedding" (e.g. Coricelli & Nagel, 2009), whereas other papers implicate control/memory brain regions, such as DLPFC (e.g. Yoshida et al., 2010). Some participants may, some of the time, discover non-mentalistic strategies to play the game, and patterns of play alone are less diagnostic than one might hope.

Thus, making progress in understanding what features or dimensions drive these regions, we believe, will again require both forward and reverse inferences. Cognitive or computational theories should suggest possible dimensions of ToM inferences that may be reflected in "more" activity in ToM regions; but at the same time, the dimensions that do, and do not, modulate the magnitude of response in ToM brain regions may provide important clues for developing theories of what these brain regions are actually doing.

Patterns within theory of mind regions

Finally, as well as looking at changes in overall activity, a third strategy is to look for divisions of functional responses across the neural populations within each region. Two relatively novel methods for

analyzing fMRI data may allow neuroscientists to look inside ToM regions. Distinctions between neural subpopulations within regions may provide clues to how these regions function.

The first method is repetition-suppression, also called functional adaptation. Repetition-suppression analyses take advantage of the observation that after processing a stimulus or task once, activity in a neuron or brain region in response to an identical stimulus or task is suppressed, or adapted (Grill-Spector, Henson, & Martin, 2006). By manipulating the features of the repeated stimulus, so that some are identical and some are different from the original stimulus, it is possible to ask what counts as the same stimulus for a particular brain region (Kourtzi & Kanwisher, 2001). If the repeated stimulus is effectively the same with regard to the features represented by the brain region, then the region's response will be suppressed or adapted. On the other hand, if a feature that is represented in the brain region has been sufficiently modified, the brain region's response will "recover" from adaptation.

This method has been used frequently and effectively to study visual representations of objects and places (Poldrack, 2006b). To date, only one study has taken advantage of this approach to test hypotheses about ToM. Jenkins, Macrae, & Mitchell (2008) asked whether attributing a preference (e.g. "enjoys winter sports") to oneself, a similar stranger, or a dissimilar stranger, depend on the same neural subpopulations within mPFC. They found that thinking about another, similar other person after thinking about the self led to repetition suppression, while thinking about the self and then a dissimilar other led to recovery from adaption. These results support the hypothesis that the mPFC represents other minds in terms of their similarity to the self.

Currently, we are using a similar strategy to investigate the components of mental state attributions. People read short stories in which a key mental state was repeated twice, with some elements changed. After the first mental state sentence (e.g. "Megan thinks that Julie is being too flirty"), the repetition either changed the agent (e.g. "Gina thinks"), the attitude verb (e.g. "Megan **worries** that"), the content (e.g. "Megan thinks that Julie should be more flirty"), all three, or none of these elements. In preliminary data, we find that ToM brain regions recover from adaptation for each kind of change on its own (compared to no change), suggesting that these regions encode all three elements of a mental state attribution. Interestingly, while the mPFC shows the most recovery when the content of the mental state changes, the left temporoparietal junction (LTPJ) shows the most recovery when the attitude verb changes, and the RTPJ shows equal recovery for any kind of change. If these results hold up to further analyses, they may provide clues about the contributions of each ToM brain region to thinking about the minds of others.

The second method, multi-voxel pattern analysis (MVPA), looks for subtle, but reliable spatial patterns within a single region (or local neighborhood) of cortex. By looking at spatial separability, MVPA provides a more direct measure of the existence of functionally separable sub-populations of neurons than repetition suppression. Each "voxel" (the fMRI equivalent of a pixel) may have some, possibly very small, preference for one kind of stimulus over another due to biases in the neuronal populations toward one type of information-processing vs. another; pattern analyses measure the similarities and differences between these patterns across space (Haxby, 2001). A few studies have begun to use pattern analyses to study ToM (e.g. Gilbert et al., 2008). In one promising example, Peelen, Wiggett, & Downing (2006) found that the pattern of activation in the mPFC reliably reflected the content of another person's emotion (e.g. sad vs. angry), independent of the stimulus modality (e.g. vocal expressions vs. body posture). Recently, we found that MVPA can be used to distinguish types of mental states within the RTPJ (Koster-Hale et al., 2013). Specifically, we find that the spatial patterns of responses across voxels (but not the magnitude of response) distinguished between harms committed intentionally vs. accidentally. This distinction cannot be detected in the pattern of activity in any other ToM brain region (or in any other part of the brain).

Moreover, we find that individual differences in the neural pattern predict individual differences in moral judgment: the individuals who have the most distinct neural patterns are also those who show the greatest behavioral difference in their moral judgments of intentional vs. accidental harms. Together, these results begin to show which distinctions are represented within a region, and point toward the underlying distinctions and computations in ToM. We are very excited by this line of inquiry, and expect that both repetition suppression and multi-voxel pattern analyses will be important contributors to the next decade of neuroimaging studies of ToM.

Limits of neuroimaging

We are optimistic that there is a lot still to learn about ToM from neuroimaging. We hope that the “neuroimaging” chapter of *Understanding Other Minds*, 4th edition will be as different from this one as this one is from the Friths’ chapter in the 2nd edition. However, it is also important to be realistic. Neuroimaging is cumbersome, expensive, and fundamentally limited. Many basic questions about ToM cannot be addressed with neuroimaging. For example, a scientific theory of how humans understand other minds should address questions like: “When and why do we (spontaneously) seek to understand another’s thoughts?”, “How do we figure out the actual content of someone else’s thoughts (i.e. *what* they are thinking) from specific cues?”, “How do we choose whether or not to incorporate others’ thoughts into our own decisions and actions?”, and “Why do we care emotionally about others’ thoughts and feelings?” None of these questions have yet been approached using neuroimaging, and may pose much harder challenges than the simpler questions we have addressed so far. Contemporary neuroimaging technology does not even allow us to address many fundamental questions about the neural mechanisms of ToM. Existing tools are extremely slow and blurry, by comparison to the speed and precision of neural computation: they cannot decipher what is the input of a region, how that input is transformed, or where the output from that region is sent, during a ToM task.

If our final horizon is a complete theory of how brain regions allow us to understand other minds, we will need to make dramatic progress on (1) the “psychophysics” of ToM in adulthood, to allow precise quantitative measurements of people’s use of ToM; (2) a computational model of ToM that is sufficiently explicit to make quantitative predictions about adult judgments (e.g. Baker, Saxe, & Tenenbaum, 2011); and (3) a mechanism of how neurons and networks of neurons might implement that computational model, by sequentially transforming patterns of input into patterns of output that make different information explicit. That horizon is still far away. However, the fact that we can give a characterization of some of the boundary conditions that a successful account of ToM needs to meet is part of what makes this such an exciting time to participate in the cognitive neuroscience of understanding other minds.

References

- Adams, R., Jr., Rule, N. O., Franklin Jr, R. G., Wang, E., Stevenson, M. T., Yoshikawa, S. (2010). Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience* 22(1): 97–108.
- Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annual review of psychology* 60: 693.
- Adolphs, R. (2010). Conceptual challenges and directions for social neuroscience. *Neuron*, 65(6): 752–67.
- Aichhorn, M. et al. (2006). Do visual perspective tasks need theory of mind? *NeuroImage*, 30(3): 1059–68.
- Aichhorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: Falseness, beliefs, or attention. *Journal of Cognitive Neuroscience* 21(6): 1179–92.

- Ames, D. L., Jenkins, A. C., Banaji, M. R., & Mitchell, J. P. (2008). Taking another person's perspective increases self-referential neural processing. *Psychological Science* 19(7): 642–4.
- Apperly, I. & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review* 116(4): 953–70.
- Apperly, I. A., Samson, D., Chiavarino, C., Bickerton, W. L., & Humphreys, G. W. (2007). Testing the domain-specificity of a theory of mind deficit in brain-injured patients: Evidence for consistent performance on non-verbal, “reality-unknown” false belief and false photograph tasks. *Cognition* 103(2): 300–321.
- Apperly, I. A., Samson, D., & Humphreys, G. W. (2005). Domain-specificity and theory of mind: evaluating neuropsychological evidence. *Trends in Cognitive Sciences* 9(12): 572–7.
- Arzy, S., Thut, G., Mohr, C., Michel, C. M., & Blanke, O. (2006). Neural basis of embodiment: distinct contributions of temporoparietal junction and extrastriate body area. *Journal of Neuroscience* 26(31): 8074–81.
- Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: modeling joint belief-desire attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*, pp. 2469–74.
- Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: The case of information overload. *Journal of Personality and Social Psychology* 49(5): 1129.
- Baron-Cohen, S., O'Riordan, M., Jones, R., Stone, V., & Plaisted, K. (1999). A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome. *Journal of Autism and Developmental Disorders* 29, 407–18.
- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders* 34(2): 163–75.
- 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* 42(2): 241–51.
- Bedny, M., Pascual-Leone, A., & Saxe, R. R. (2009). Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences of the United States of America* 106(27): 11312–17.
- Bhatt, M., Lohrenz, T., Camerer, C. F., & Montague, P. R. (2010). Neural signatures of strategic types in a two-person bargaining game. In *Proceedings of the National Academy of Sciences* 107(46): 19720–5
- Blanke, O., & Arzy, S. (2005). The out-of-body experience: disturbed self-processing at the temporo-parietal junction. *Neuroscientist* 11(1): 16–24.
- Blanke, O., Mohr, C., Michel, C. M., Pascual-Leone, A., Brugger, P., Seeck, M., et al. (2005). Linking out-of-body experience and self processing to mental own-body imagery at the temporoparietal junction. *Journal of Neuroscience* 25(3): 550–7.
- Bloom, P., & German, T. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77(1): B25–31.
- Bruneau E., Dufour N., & Saxe R. (2012) Social cognition in members of conflict groups: behavioural and neural responses in Arabs, Israelis and South Americans to each other's misfortunes. *Philosophical Transactions of the Royal Society, London, B Biological Sciences* 367(1589): 717–30.
- Bruneau, E. G., Pluta, A., & Saxe, R. (2011). Distinct roles of the “shared pain” and “theory of mind” networks in processing others’ emotional suffering. *Neuropsychologia* pp.1–13.
- Bruneau, E. G., & Saxe, R. (2010). Attitudes toward the outgroup are predicted by activity in the precuneus in Arabs and Israelis. *NeuroImage* 52(4): 1704–11.
- Carrington, S. J., & Bailey, A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping* 30(8): 2313–35.

- Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage* 12(3): 314–25.
- Cloutier, J., Gabrieli, J. D. E., O'Young, D., & Ambady, N. (2011). An fMRI study of violations of social expectations: When people are not who we expect them to be. *NeuroImage* 57(2): 583–8.
- Cohen, S. B., Wheelwright, S., & Hill, J. (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*, 42(2): 241–51.
- Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience* 3: 292–7.
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Review Neuroscience* 3: 201–15.
- Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron* 58(3): 306–24.
- Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences* 106(23): 9163–8.
- Costa, A., Torriero, S., & Oliveri, M. (2008). Prefrontal and temporo-parietal involvement in taking others' perspective: TMS evidence. *Behavioural Neurology* 19(1–2): 71–4.
- Costafreda, S. G., Brammer, M. J., David, A. S., & Fu, C. H. (2008). Predictors of amygdala activation during the processing of emotional stimuli: A meta-analysis of 385 PET and fMRI studies. *Brain Research Reviews* 58(1): 57–70.
- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108(2): 353–80.
- D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., et al. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Cognitive Neuroscience* 19(6): 935–944.
- Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *The Neuroscientist*, 13(6): 580–93.
- Decety, J., Michalska, K. J., & Akitsuki, Y. (2008). Who caused the pain? An fMRI investigation of empathy and intentionality in children. *Neuropsychologia* 46(11): 2607–14.
- Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage* 55(2): 705–12.
- Döhnell, K., Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., & Sommer, M. (2012). Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *NeuroImage* 60(3): 1652.
- Dufour, N., Redcay, E., Young, L., Mavros, P., Moran, J., Triantafyllou, C., Gabrieli, J., & Saxe, R. (2012). What explains variability in brain regions associated with Theory of Mind in a large sample of neurotypical adults and adults with ASD? In N. Miyake, D. Peebles, & R. P. Cooper (Eds), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 312–17). Austin: Cognitive Science Society.
- Fedorenko, E., & Kanwisher, N. (2009). Neuroimaging of language: Why hasn't a clearer picture emerged? *Language and Linguistics Compass* 3(4): 839–65.
- Fletcher, P., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of “theory of mind” in story comprehension. *Cognition* 57(2): 109–28.
- Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature* 12(9): 1187–96.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping* 2: 189–210.

- Frith, C. D., & Frith, U. (2000). The physiological basis of theory of mind. In: S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds), *Understanding Other Minds: Perspective from Developmental Social Neuroscience* (pp. 335–56). Oxford: Oxford University Press.
- Frith, C., & Frith, U. (2012). Social neuroscience. *Annual Review of Psychology* **63**: 287–313.
- Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of “theory of mind”. *Trends in Cognitive Sciences* **7**, 77–83.
- Gallagher, H., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of “theory of mind” in verbal and nonverbal tasks. *Neuropsychologia* **38**(1): 11–21.
- Georgopoulos, A., Schwartz, A., & Kettner, R. (1986). Neuronal population coding of movement direction. *Science* **233**(4771): 1416–19.
- Gilbert, S. J., Meuwese, J. D., Towgood, K. J., Frith, C. D., & Burgess, P. W. (2009). Abnormal functional specialization within medial prefrontal cortex in high-functioning autism: a multi-voxel similarity analysis. *Brain* **132**(4): 869–78.
- Gobbini, M., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience* **19**(11): 1803–14.
- Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *NeuroReport* **6**: 1741–6.
- Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.
- Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences* **10**(1): 14–23.
- Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child Development* **83**: 1853–68.
- Hamilton, D. L. (1988). Causal attribution viewed from an information-processing perspective. In: D. Bar-Tal & A. W. Kruglanski (Eds), *The Social Psychology of Knowledge* (pp. 359–85). Cambridge: Cambridge University Press.
- Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review* **103**(2): 336.
- Haxby, J. V. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**(5539): 2425–30.
- Hayashi, A., Abe, N., Ueno, A., Shigemune, Y., Mori, E., Tashiro, M., & Fujii, T. (2010). Neural correlates of forgiveness for moral transgressions involving deception. *Brain Research* **1332**: 90–9.
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology* **57**(2): 243–59.
- Hein, G., & Knight, R. T. (2008). Superior temporal sulcus—it’s my area: or is it? *Journal of Cognitive Neuroscience* **20**(12): 2125–36.
- Higgins, E. T., & Bargh, J. A. (1987). Social cognition and social perception. *Annual Review of Psychology* **38**(1): 369–425.
- Jacques, P., Conway, M. A., Lowder, M. W., & Cabeza, R. (2011). Watching my mind unfold versus yours: An fMRI study using a novel camera technology to examine neural differences in self-projection of self vs. other perspectives. *Journal of Cognitive Neuroscience* **23**(6): 1275–84.
- Jenkins, A., Macrae, C., & Mitchell, J. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences* **105**(11): 4507.
- Jenkins, A., & Mitchell, J. (2010). Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex* **20**(2): 404–410.
- Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., et al. (2010). Dissociating cognitive from affective theory of mind: a TMS study. *Cortex*, **46**(6): 769–80.
- Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* **8**(5): 679–85.

- Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. In *Proceedings of the National Academy of Sciences* **107**(25): 11163–70.
- Kobayashi, C., Glover, G. H., & Temple, E. (2007). Children's and adults' neural bases of verbal and nonverbal "theory of mind". *Neuropsychologia* **45**: 1522–32.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* **446**(7138): 908–11.
- Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipito-temporal cortex. *Neuron* **74**(6): 1114–24.
- Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences* **110**(14): 5648–5653.
- Koster-Hale, J., & Saxe, R. R. (2011). Theory of Mind brain regions are sensitive to the content, not the structural complexity, of belief attributions. In: L. Carlson, C. Hoelscher, & T. F. Shipley (Eds), *Proceedings of the 33rd Annual Cognitive Science Society Conference*, pp. 3356–61.
- Kourtzi, Z. & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science* **293**(5534): 1506.
- Knight, R. T. (2007). Neural networks debunk phrenology. *Science* **316**(5831): 1578–9.
- Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage* **38**(4): 649–62.
- Kriegeskorte, N., Goebel, R., & Bandettini, P. A. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences USA* **103**(10), 3863–8.
- Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience* **30**(41): 13906–15.
- Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience* **19**(1): 42–58.
- Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004). Measuring empathy: reliability and validity of the empathy quotient. *Psychological Medicine*, **34**(5): 911–24.
- Leslie, A. M., & Frith, U. (1990). Prospects for a cognitive neuropsychology of autism: Hobson's choice. *Psychological Review* **97**(1): 122–31.
- Liu, D., Sabbagh, M., A., Gehring, W. J., & Wellman, H. M. (2004). Decoupling beliefs from reality in the brain: an ERP study of theory of mind. *Neuroreport* **15**(6): 991–5.
- Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, **453**(7197): 869–78.
- Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., et al. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience*, **22**(7): 1623–35.
- Mars, R. B., Sallet, J., Schüffegen, U., Jbabdi, S., Toni, I., & Rushworth, M. F. (2012). Connectivity-based subdivisions of the human right "temporoparietal junction area": Evidence for different areas participating in different cortical networks. *Cerebral Cortex* **22**(8): 1894–903.
- Mason, R. A., & Just, M. A. (2010). Differentiable cortical networks for inferences concerning people's intentions vs. physical causality. *Human Brain Mapping*, **32**(2): 313–29.
- Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**: 167–202.
- Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, **18**(2): 262–71.
- Mitchell, J. P., Banaji, M. R., & MacRae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience* **17**(8): 1306–15.
- Mitchell, J., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, **50**(4): 655–63.

- Moran, J. M., Lee, S. M., & Gabrieli, J. D. E. (2011a). Dissociable neural systems supporting knowledge about human character and appearance in ourselves and others. *Journal of Cognitive Neuroscience*, 23(9): 2222–30.
- Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O’Young, D., Mavros, P. L., & Gabrieli, J. D. (2011b). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, 108(7): 2688–92.
- Moriguchi, Y., Ohnishi, T., Lane, R. D., Maeda, M., Mori, T., Nemoto, K., et al. (2006). Impaired self-awareness and theory of mind: An fMRI study of mentalizing in alexithymia. *NeuroImage* 32(3): 1472–82.
- Onishi, K., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719): 255.
- Peelen, M. V., Wiggett, A. J., & Downing, P. E. (2006). Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron*, 49(6): 815–22.
- Pelphrey, K. A., Mitchell, T. V., McKeown, M. J., Goldstein, J., Allison, T., & McCarthy, G. (2003). Brain activity evoked by the perception of human walking: Controlling for meaningful coherent motion. *Journal of Neuroscience*, 23, 6819–25.
- Pelphrey, K. A., Morris, J. P., Michelich, C. R., Allison, T., & McCarthy, G. (2005). Functional anatomy of biological motion perception in posterior temporal cortex: An fMRI study of eye, mouth and hand movements. *Cerebral Cortex*, 15(12): 1866–76.
- Pelphrey, K. A., & Morris, J. P. (2006). Brain mechanisms for interpreting the actions of others from biological-motion cues. *Current Directions in Psychological Science*, 15(3): 136.
- Perner, J. (1991). *Understanding the Representational Mind*. Cambridge: MIT Press.
- Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience* 1(3–4): 245–58.
- Platek, S., Keenan, J. P., Gallup Jr, G. G., & Mohamed, F. B. (2004). Where am I? The neurological correlates of self and other. *Cognitive Brain Research* 19(2): 114–22.
- Poldrack, R. A. (2006a). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* 10(2): 59–63.
- Poldrack, R. A. (2006b). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience* 2(1): 67–70.
- Rabin, J. S., Gilboa, A., Stuss, D. T., Mar, R. A., & Rosenbaum, R. S. (2010). Common and unique neural correlates of autobiographical memory and theory of mind. *Journal of Cognitive Neuroscience* 22(6): 1095–111.
- Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology* 33(1): 12.
- Samson, D., Apperly, I. A., & Humphreys, G. W. (2007). Error analyses reveal contrasting deficits in “theory of mind”: Neuropsychological evidence from a 3-option false belief task. *Neuropsychologia* 45(11): 2561–9.
- Saxe R. (In press). The new puzzle of theory of mind development. In: M. R. Banaji & S. A. Gelman (Eds), *The Development of Navigating the Social Cognition. World: What infants, children, and other species can teach us*. New York: Oxford University Press.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind.” *NeuroImage* 19(4): 1835–42.
- Saxe, R., Moran, J. M., Scholz, J., & Gabrieli, J. (2006a). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience* 1(3): 229–34.
- Saxe, R., & Offen, S. (2010). Seeing ourselves: What vision can teach us about metacognition. In: G. Dimaggio, & P. H. Lysaker (Eds), *Metacognition and Severe Adult Mental Disorders: From basic research to treatment* (pp. 13–29). New York: Taylor & Francis.

- Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science* 17(8): 692–9.
- Saxe, R. R., Schulz, L. E., & Jiang, Y. V. (2006b). Reading minds vs. following rules: dissociating theory of mind and executive control in the brain. *Social Neuroscience*, 1(3–4): 284–98.
- Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia* 43(10): 1391–9.
- Saxe, R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Development*, 80(4): 1197–209.
- Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS ONE* 4(3): e4869–e4869.
- Schnell, K., Bluschke, S., Konradt, B., & Walter, H. (2011). Functional relations of empathy and mentalizing: An fMRI study on the neural basis of cognitive empathy. *NeuroImage* 54(2): 1743–54.
- Serences, J. T., Shomstein, S., Leber, A. B., Golay, X., Egeth, H. E., & Yantis, S. (2005). Coordination of voluntary and stimulus-driven attentional control in human cortex. *Psychological Science* 16:114–22.
- Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. *Neuropsychologia* 45(13): 3054–67.
- Shamay-Tsoory, S., Tomer, R., Berger, B. D., Goldsher, D., & Aharon-Peretz, J. (2005). Impaired affective theory of mind is associated with right ventromedial prefrontal damage. *Cognitive and Behavioral Neurology* 18(1): 55–67.
- Slessor, G., Phillips, L. H., & Bull, R. (2007). Exploring the specificity of age-related differences in theory of mind tasks. *Psychology and Aging*, 22(3): 639–43.
- Sommer, M., Döhnelt, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. *NeuroImage*, 35(3): 1378–84.
- Sommer, M., Rothmayr, C., Döhnelt, K., Meinhardt, J., Schwerdtner, J., Sodian, B., & Hajak, G. (2010). How should I decide? The neural correlates of everyday moral reasoning. *Neuropsychologia*, 48(7): 2018–26.
- Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science* 18(7): 587–92.
- Spiers, H. J., & Maguire, E. A. (2006). Thoughts, behaviour, and brain dynamics during navigation in the real world. *NeuroImage*, 31(4): 1826–40.
- Spunt, R. P., & Lieberman, M. D. (2012). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. *NeuroImage* 59(3), 3050.
- Spunt, R., Satpute, A. B., & Lieberman, M. (2011). Identifying the what, why, and how of an observed action: an fMRI study of mentalizing and mechanizing during action observation. *Journal of Cognitive Neuroscience* 23(1): 63–74.
- Tamir, D. I. & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences*, 107(24): 10827.
- Tsakiris, M., Costantini, M., & Haggard, P. (2008). The role of the right temporo-parietal junction in maintaining a coherent sense of one's body. *Neuropsychologia* 46(12): 3014–18.
- Uttal, W. R. (2011). *Mind and Brain: A Critical Appraisal of Cognitive Neuroscience*. Cambridge: MIT Press.
- Van Overwalle, F. (2008). Social cognition and the brain: A meta-analysis. *Human Brain Mapping* 30(3): 829–58.
- Vogele, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., ... & Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage* 14(1): 170–81.
- Vogele, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience* 16, 817–27.

- Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain:: an event-related fMRI study. *Neuron*, 30(3): 829–41.
- Wagner, D. D., Kelley, W. M., & Heatherton, T. F. (2011). Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. *Cerebral Cortex*, 21(12): 2788–96.
- Walter, H., Schnell, K., Erk, S., Arnold, C., Kirsch, P., Esslinger, C., et al. (2010). Effects of a genome-wide supported psychosis risk variant on neural activation during a theory-of-mind task. *Molecular Psychiatry* 16(4): 462–70.
- Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3): 655–84.
- Whitfield-Gabrieli, S., Moran, J. M., Nieto-Castañón, A., Triantafyllou, C., Saxe, R., & Gabrieli, J. D. (2011). Associations and dissociations between default and self-reference networks in the human brain. *NeuroImage*, 55(1): 225–32.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13(1): 103–28.
- Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism* 12(6): 900–18.
- Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *Journal of Neuroscience* 30(32): 10744–51.
- Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010a). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, 107(15): 6753–8.
- Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences* 104(20): 8235–40.
- Young, L., Dodell-Feder, D., & Saxe, R. (2010b). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, 48(9): 2658–64.
- Young, L., Nichols, S., & Saxe, R. (2010c). Investigating the neural and cognitive basis of moral luck: it's not what you do but what you know. *Review of Philosophy and Psychology*, 1(3): 333–49.
- Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, 40(4): 1912–20.
- Young, L., & Saxe, R. (2009a). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience* 21(7): 1396–405.
- Young, L., & Saxe, R. (2009b). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia* 47(10): 2065–72.
- Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and. *Cognition* 35: 41–68.
- Zaitchik, D., Koff, E., Brownell, H., Winner, E., & Albert, M. (2006). Inference of beliefs and emotions in patients with Alzheimer. *Neuropsychology* 20, 11–20.
- Zaitchik, D., Walker, C., Miller, S., LaViolette, P., Feczko, E., & Dickerson, B. C. (2010). Mental state attribution and the temporoparietal junction: An fMRI study comparing belief, emotion, and perception. *Neuropsychologia* 48(9): 2528–36.

Theory of mind: Insights from patients with acquired brain damage

Dana Samson and Caroline Michel

There is a general agreement that one way in which we are able to understand other people's minds is by using knowledge and processes by which we impute invisible mental states such as beliefs, intentions or emotions in order to explain and predict people's behavior, an ability referred to as *Theory of Mind* (ToM, Premack & Woodruff, 1978). Yet despite this consensus, it remains unclear what cognitive and neural mechanisms underlie this ability.

Some evidence suggests that ToM relies on effortful processes. For example, children start to reason accurately and explicitly about complex mental states, such as beliefs, at the same time as they undergo important developments in their language and executive function abilities (Carlson & Moses, 2001; Villiers & Pyers, 2002). Furthermore, adults do not seem to always engage automatically in belief reasoning (Apperly, Riggs, Simpson, Chiavarino, & Samson, 2006a) and they often fail to take someone else's perspective into account (Keysar, Barr, Balin, & Brauner, 2000), especially when they are under cognitive load or time pressure (Epley, Keysar, Van Boven, & Gilovich, 2004).

There is also evidence that ToM is *not a unitary function*. For example, neuroimaging studies have shown that reasoning about other people's mental states activates a network of distinct brain areas (for a meta-analysis, see Van Overwalle, 2009; see also Chapter 9). Furthermore, lesions to different brain areas can lead to ToM impairments. The first reports of ToM impairments were observed following acquired brain lesions to the right hemisphere (Happé, Brownell, & Winner, 1999; Siegal, Carrington, & Radel, 1996; Surian & Siegal, 2001) and the frontal lobes (Lough & Hodges, 2002; Stone, Baron-Cohen, & Knight, 1998; Stuss, Gallup, & Alexander, 2001), but there is now also evidence of impairments following damage to the temporo-parietal junction (Samson, Apperly, Chiavarino, & Humphreys, 2004), and subcortical structures such as the amygdala (Stone, Baron-Cohen, Calder, Young, & Keane, 2003) and the basal ganglia (Bodden, Dodel, & Kalbe, 2010). This diversity of lesion localization is consistent with the idea that ToM is sustained by a distributed network of brain regions, and it is very likely that different brain regions play different roles.

While we have good reasons to expect that ToM relies, at least partly, on effortful processes and that it should not necessarily be construed as a unitary function, we still lack a detailed account of the cognitive and neural architecture of ToM. The aim of this chapter is to show how the study of the patterns of association and dissociation of deficits in patients with acquired brain lesions can help identify the building blocks of ToM and specify the nature of these building blocks in relation to high-order cognition functions such as executive function and language. Indeed, if ToM can be broken down in sub-processes that are relatively independent one from another at the functional and neural level, then we can expect that an acquired brain lesion may selectively affect one type of

ToM sub-processes while leaving the others unaffected. Furthermore, if some ToM sub-processes rely on high-order functions such as executive function and language, we may expect that a deficit affecting the putative critical executive or language processes should impact on the efficiency with which these ToM sub-processes are operating.

In the following sections, we show the contribution of neuropsychological studies in highlighting what may be some of the ToM building blocks and what may be the nature of these building blocks at the cognitive and neural level. In doing so, we will mainly focus on ToM knowledge and processes which allow us to infer **cognitive**, rather than affective mental states, the latter being the focus of a different chapter (see Chapter 11).

Inhibiting one's own perspective

Other people's desires, emotions, or knowledge are often different to our own, and it is thus crucial to be able to put our own perspective aside when we try to understand or predict other people's behaviors. This is illustrated in one of the most famous examples of ToM scenarios: the "Sally and Anne" story (Baron-Cohen, Leslie, & Frith, 1985). In that story, Sally puts her marble into her

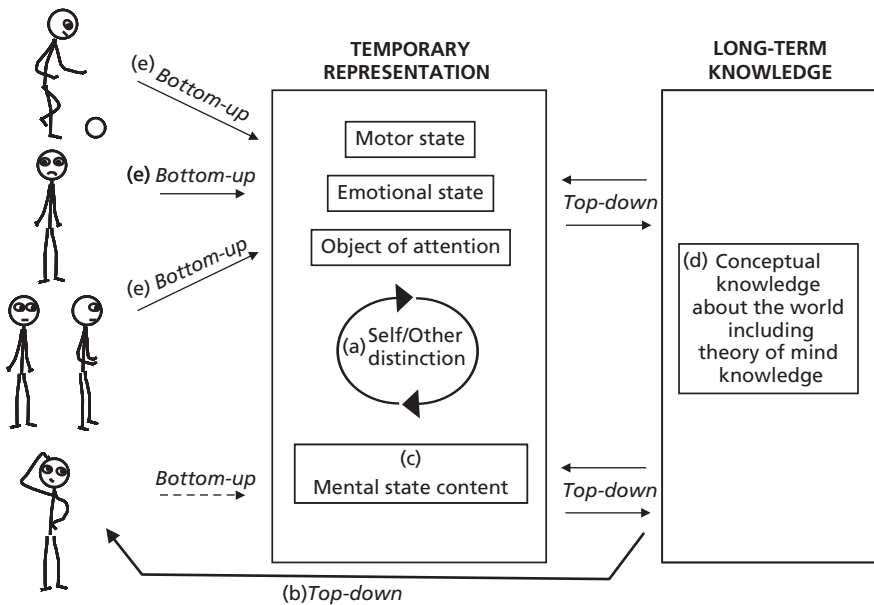


Figure 10.1 A schematic representation of the hypothesized components underlying our ability to understand other people's minds (adapted from Samson, 2009). Components (a–d) are traditionally associated with ToM and include: (a) processes allowing us to distinguish our own mental states from that of others, including mechanisms inhibiting our own perspective; (b) processes allowing us to monitor the environment in order to track relevant cues to infer a content to other peoples' mental states; (c) a short-term memory system allowing us to hold in mind the inferred content of mental states; and (d) a long-term memory system storing ToM knowledge. Components (e) are low-level and bottom-up processes which can bypass and/or feed into ToM components.

Adapted from Samson, D. Reading other people's mind: insights from neuropsychology. *Journal of Neuropsychology*, 3(1): 3–16 © 2009, John Wiley and Sons, with permission.

basket and goes out for a walk. While she is away, Anne takes the marble out of the basket and puts it into her box. Sally then comes back and wants to play with her marble. If asked where Sally will first look for her marble, most adults (and already most children from the age of 5; Wimmer & Perner, 1983) will say that Sally will first open her basket, despite the fact that they know that the marble is not in there, and despite the fact that they themselves would have opened the box, rather than the basket if asked to get the marble. Thus, in order to correctly predict what Sally will do, we have to put our own perspective aside in order to realize that Sally has a different view about the location of the object.

Neuropsychological studies have shown that the ability to put one's own perspective aside can be selectively affected by acquired brain damage, suggesting that some cognitive and neural processes are specifically dedicated to the inhibition of our own perspective and not to other aspects of perspective taking (see (a) in Figure 10.1). Direct evidence for this comes from a single case study of a patient, WBA, who sustained right lateral prefrontal brain lesions following a stroke (Samson, Apperly, Kathirgamanathan, & Humphreys, 2005). The performance of patient WBA was compared across two non-verbal and video-based false belief tasks in which the demands in terms of self-perspective inhibition were manipulated (see Figure 10.2). In both tasks and on false belief trials, a protagonist was not aware that the location of an object had changed. In the task with high self-perspective inhibition demands (reality-known condition), and similarly to classic false belief scenarios, the patient was shown the new location of the target object and had to resist interference from that knowledge in order to infer that the protagonist in the scenario had a false belief. Patient WBA was unable to ascribe a false belief to the protagonist in those conditions (he scored 1/12 correct). In contrast, in the task with low self-perspective inhibition demands (reality-unknown condition), the patient wasn't shown the precise location of the object before or after the change, but was simply made aware that the object location had changed. Here, knowledge of the location of the object could not interfere with the realization that the protagonist's belief was false (see Figure 10.2). In that latter case, patient WBA was able to ascribe a false belief to the protagonist (he scored 11/12 correct). Thus, patient WBA was only impaired at ascribing a false belief when he knew the true state of the world.

Further evidence supporting the claim that patient WBA suffered from a selective deficit in inhibiting his own perspective comes from another modified false belief task in which three types of response options were given. The patient watched a protagonist witnessing what another protagonist put in a branded box (e.g. protagonist B watched protagonist A put an apple in an empty cornflakes box). On false belief trials, protagonist B then left the scene, and in his absence, protagonist A changed the content of the box (e.g. protagonist A removed the apple from the cornflakes box and put a wooden spoon in the cornflakes box). Protagonist B then came back to the scene, and the patient was asked what protagonist B thought is inside the box. Three options were given: (1) protagonist B thinks there is an apple in the box (i.e. the response expected if the patient correctly inferred the protagonist's false belief), (2) protagonist B thinks there is a wooden spoon in the box (i.e. the response expected if the patient transposed his own knowledge of the real state of the world), or (3) protagonist B thinks there are cornflakes in the box (i.e. the response expected if the patient used a simplified strategy that bypasses the integration of the previous sequences of events in order to infer the belief content). In this task, patient WBA only scored 2/16 correct on false belief trials, and all his errors consisted in wrongly transposing his own perspective to that of the other person (Samson, Apperly, & Humphreys, 2007).

Several additional observations have helped characterize the processes underlying the ability to inhibit one's own perspective. First, patient WBA's difficulties in inhibiting his own perspective were not only apparent in belief reasoning tasks, but were also apparent when he was asked to infer

someone else's visual experience, desire, or emotion (Samson et al., 2005). This suggests that the same self-perspective inhibition processes may be involved irrespective of the nature of the mental state to reason about.

Secondly, WBA's brain lesions also affected his performance on standard executive function tasks, including tasks measuring general inhibition abilities. However, recent findings suggest that not all patients with inhibitory control difficulties will show a selective deficit in self-perspective inhibition (Houthuys, 2010). Patients with frontal lesions can show different types of inhibitory control difficulties. They can have difficulties inhibiting salient, but irrelevant external information, which can lead to a high distractibility and an inability to focus on a task goal. Alternatively, they can have difficulties inhibiting internally generated irrelevant thoughts, which can lead to perseveration of thoughts and actions as well as an inability to change their thoughts or actions in response to external cues (Burgess, Dumontheil, & Gilbert, 2007). It seems that it is the latter type of patients who suffer from a deficit in self-perspective inhibition and not the former type (Houthuys, 2010). Thus, we could speculate that the inhibitory processes involved in resisting interference from our own perspective are not the same as just any inhibitory processes, but that they may be the same as those that control internally generated thoughts. This would need to be confirmed by other studies.

Finally, patient WBA's lesion location points to the right lateral prefrontal cortex as part of the neural substrate of our ability to inhibit our own perspective. This is consistent with a recent neuroimaging study which also points to that region as being involved in the inhibition of one's own perspective (van der Meer, Groenewold, Nolen, Pijnenborg, & Aleman, 2011).

The evidence reported here from patients with acquired brain damage can have a wider impact in understanding the origin of ToM difficulties in other populations. For example, by using a similar approach of contrasting ToM tasks with high and low demands in self-perspective inhibition, it has been shown that ageing disproportionately affects the ability to inhibit one's own perspective (Bailey & Henry, 2008).

Monitoring the environment

While one may be able to successfully put aside one's own perspective (realizing that the other person may have a different mental state), this ability alone would still not be sufficient to infer the specific content of the other person's perspective. We also need to select, monitor, and integrate the relevant cues in the specific situation at hand in order to provide the appropriate inputs for reasoning about the other person's mental state content (see (b) in Figure 10.1). For example, what is the type of relevant information we need to take into account to predict where Sally is going to look for her marble? How far back in the past do we need to go to find the relevant information? Neuropsychological studies indicate that some processes may be specifically dedicated to the monitoring of the environment, rather than to other aspects of ToM processing.

Evidence for a role of environment monitoring comes from another case study of a patient, PF, who suffered from a stroke affecting the left temporo-parietal areas of the brain (Samson et al., 2007, 2004). Unlike patient WBA, patient PF was not more impaired in the false belief task that placed the highest demands on self-perspective inhibition. On the contrary, she performed better on the task with high self-perspective inhibition demands (scoring 11/12 correct) than the task with low self-perspective inhibition demands (scoring 2/12 correct). One explanation for this profile of performance is that PF was sensitive to how directly the instructions invited her to consider the other person's perspective. Indeed, besides the varying demands in self-perspective inhibition, the reality-known and reality-unknown false belief tasks also differed in their task instructions



Figure 10.2 Sequences of events in the false belief trials of the reality-known and the reality-unknown belief reasoning tasks (Apperly et al., 2004; Samson et al., 2004, 2005). In both tasks, false belief trials are mixed with true belief and control trials. (a) The reality-known false belief task matches the classic false belief paradigm. Participants are asked to predict which box the woman will open first to find the green object. On the false belief trials, the woman is unaware of the change of location and participants know the new location of the object. Thus, to infer that the woman will open the wrong box, participants need to resist interference from their own knowledge of the object's location (high self-perspective inhibition condition). (b) The reality-unknown false belief task was adapted from a task used with non-human animals (Call & Tomasello, 1999). Participants are asked to locate the green object and are told that the woman will try to help them find the object by using a pink marker. At the beginning of each trial, only the woman (and not the participant) is shown the content of the boxes. On false belief trials, the woman is unaware of the change of location and, participants know that there has been a change of location, but do not know where the object is located. Thus, to infer that the woman points to the wrong box, participants do not have to deal with the interference of their own knowledge of the object's location (low self-perspective inhibition condition).

(see Figure 10.2). In the reality-known condition, i.e. the task in which PF scored well, the instructions explicitly stated that she had to predict what the other person would do. In contrast, in the reality-unknown condition, i.e. the task in which PF performed poorly, PF was asked to locate a target object, and nothing in the instructions reminded her that the other person's perspective may have been relevant to locate the object. Patient PF's difficulties may have thus arisen from not spontaneously tracking the other person's belief.

Furthermore, in the three-option false belief task described earlier, only 20% of PF's errors consisted in transposing her own perspective, unlike patient WBA, for whom 100% of his errors were egocentric errors (Samson et al., 2007). PF's other errors consisted in choosing the most likely content of the branded box, suggesting that PF did not take into account the previous sequences of events, and may have fallen back on a simplified mentalizing strategy of imputing as belief content what the person may have inferred from simply "looking" at the box. Two further observations of PF's performance across the different false belief tasks are worth noting: (i) her errors were mostly made in the first half of the tasks and (ii) her errors were not confined to false belief trials; they also appeared on true belief trials (Samson et al., 2007). It appeared as if the patient did not realize which cues would be relevant, and only found them out through trial and error. For example, over the course of trials, PF seemed to have realized that when the protagonist in the scenario left

the scene, s/he didn't know that the state of the world had changed. If on the true belief trials, the protagonist also left the room, but there was no change of the state of the world, or the person came back in the room on time to see the change, patient PF then incorrectly attributed a lack of knowledge or false belief. Thus, patient PF seemed to over-generalize the consequences of the protagonist leaving the room without integrating this information with other clues (e.g. what happened when the protagonist was away; when did the protagonist leave or come back in relation to when the changes were made in the room, etc.). Interestingly, such over-generalizations have also been reported in young children (Sodian & Thoermer, 2008). Collectively, the profile of performance of patient PF indicated that her difficulties were not linked to a deficit in self-perspective inhibition, but were rather related to a deficit in spontaneously tracking other people's mental states and/or monitoring the environment to find cues to the other person's mental state content.

Interestingly, patient PF's difficulties observed in the false belief tasks extended to a carefully matched false photograph task, where she had to infer that a photograph misrepresented the real state of the world (Apperly, Samson, Chiavarino, Bickerton, & Humphreys, 2007). This association of deficit is consistent with neuroimaging studies in healthy adults which have shown that the left temporo-parietal junction (damaged in the case of patient PF) is activated not only when healthy adults reason about false beliefs, but also when they reason about non-social representations of the world such as false signs (for example, a sign indicating the direction of a specific place can be displaced so that it indicates the wrong direction; Aichhorn, Perner, Weiss, Kronbichler, Staffen, & Ladurner, 2009). This is perhaps not surprising as even for non-social representations of the world such as photographs, it is necessary to pay attention to the correct clues to infer the content of the representation (the content of a photograph will depend, for example, on the time point at which the photograph was taken, the angle of the shot, etc.).

In sum, the pattern of responses of patient PF indicates that self-perspective inhibition is not the sole source of processing core to belief reasoning and that processes are also required to monitor the environment. It is possible that an impairment of these latter processes is at the origin of ToM difficulties in other disorders than acquired brain lesions to the temporo-parietal junction.

Constructing and holding a temporary representation in mind

Inferring someone else's mental state requires the ability to represent temporarily in memory the content of the other person's mental states (see (c) in Figure 10.1). This can be particularly complex information in the case of mental states that are propositional attitudes, such as beliefs (e.g. "Sally thinks that the marble is in the basket") or certain forms of desires (e.g. "Sally wants that Anne puts the marble back in the basket"), requiring the need to represent and hold in mind the embedded information.

There is evidence that complex embeddings require high working memory resources. For example, it has been shown that when healthy adults perform a working memory task while simultaneously performing a belief reasoning task, the efficiency with which they infer false beliefs is significantly reduced, with the detrimental effects of the concurrent working memory task being larger for second-order (e.g. "Peter thinks that Sally thinks that the marble is in the basket") than first-order false belief inferences (e.g. "Sally thinks that the marble is in the basket," McKinnon & Moscovitch, 2007). In other words, the more perspectives are needed to be monitored and integrated, the larger the detrimental effect of the concurrent working memory task. Similarly, it has been shown that brain-damaged patients' residual working memory capacities significantly predicted their performance on a second-order, but not first-order false belief task (McDonald

& Bibby, 2005). Thus, working memory resources seem to be differentially recruited depending on the complexity of the mental state contents that we represent in our minds.

Besides the role of working memory, other authors have hypothesized that language and particularly grammar would play a crucial role in representing propositional attitudes (de Villiers & Pyers, 2002). Indeed, propositional attitudes are verbalized by means of complement clauses which reflect the embedded nature of the clause (e.g. “Sally thinks that (the marble is in the basket)”), and it is therefore possible that the same grammatical structure is necessary to represent someone else’s belief in our mind.

There is evidence that children’s ability to process embedded sentences (complement or relative clauses) correlates with their performance on a false belief task (Smith, Apperly, & White, 2003; de Villiers & Pyers, 2002), suggesting that the mastering of complex grammar may give them the means to discover the properties of propositional attitudes such as beliefs. However, this does not mean that grammar is still necessary once an individual has a fully-fledged theory of mind. On the contrary, neuropsychological studies suggest that adults with grammatical processing deficits following brain damage can still reason about false beliefs (Apperly, Samson, Carroll, Hussain, & Humphreys, 2006b; Varley & Siegal, 2000; Varley, Siegal, & Want, 2001). One of these studies documented the case of a patient, PH, who suffered a left hemisphere stroke and who, as a consequence of his lesions, had severe grammar processing difficulties (Apperly et al., 2006b). When reading a sentence, patient PH relied on the meaning of each single word to infer the meaning of the whole sentence without taking into account grammar. Thus, he would be unable to distinguish the meaning of reversible sentences such as “the man is followed by the dog” and “the man follows the dog”. Patient PH was impaired both at processing relative clause sentences (he scored 7/14 correct) and complement clause sentences (he scored 3/12 correct, Apperly et al., 2006b). However, he had no difficulties in non-verbal false belief reasoning tasks, scoring between 10/12 and 12/12 correct on three different tasks which required inferring that someone else has a false belief (first-order false belief tasks) and he even scored 12/12 in a non-verbal second-order false belief task. These results indicate that although the mastering of complex grammar may help children reasoning about beliefs, grammar does not seem to play a fundamental role anymore in adulthood, not even for complex mental state contents such as second-order beliefs. Thus, so far it remains unclear in what format we hold propositional attitudes in working memory.

Theory of mind long-term knowledge

Making sense of other people’s minds requires not only a set of processes that allow us to infer other people’s mental states, but also long-term semantic knowledge about mental states that can be used to guide the inferential processes (see (d) in Figure 10.1). Although nobody would deny the existence of long-term knowledge about mental states, this “knowledge” component of the ToM architecture has so far received little attention in the literature.

Some authors have construed ToM knowledge as including abstract general rules or laws about the mind (i.e. ToM knowledge is seen as a folk theory about how the mind works; e.g. Gopnik, 2003; Gopnik & Wellman, 1992). These rules or laws would specify the relation between mental states and objects or events in the world (e.g. given that an object is within a viewer’s line of sight, the viewer will see it), the relation between mental states and behaviors (e.g. people’s actions are determined by their representation of the world rather than by the world itself), the relations amongst mental states (e.g. perceptions lead to beliefs), as well as other properties (e.g. mental states have a tendency to change, mental states differ between individuals). While such a conceptualization of ToM as a “theory” has been questioned (see for example Gordon, 1986), it is more widely accepted

that ToM knowledge includes a set of concepts referring to different types of mental states, with each concept differing from others according to specific semantic features. For example, concepts referring to epistemic mental states can vary according to the level of certitude that they represent (e.g. knowing, thinking, or guessing). Adults' rich repertoire of mental state terms testifies that we code in memory fine-grained conceptual distinctions between mental states, and it has been proposed that it is through the exposure to and the acquisition of a rich lexical repertoire that children discover the distinctive properties of ToM concepts (see for example, Peterson & Siegal, 2000). The question remains to what extent such a rich lexical repertoire and/or the associated fine-grained distinctions at the conceptual level are necessary to infer other people's mental states.

In principle, neuropsychological studies offer the possibility to investigate this question by studying how the loss of such a rich repertoire and/or the associated fine-grained conceptual distinctions impacts on a patient's ability to infer other people's mental states. In line with this approach, one study tested a patient, CM, who suffered from semantic dementia and who showed a massive atrophy of the left temporal pole (Michel, Dricot, Lhommel, Grandin, Ivanoiu, Pillon et al., 2011). The patient showed a severe impairment in tasks probing his general semantic knowledge about the world (e.g. living and non-living entities, abstract and concrete entities), and more relevant to our question of interest, the patient had a severely impoverished ability to use and understand mental state terms. For example, when asked to judge which of two words was closest in meaning to a target word, CM only scored 24/48 correct for mental state words (a score significantly below the level of performance of matched control participants who all scored above 42/48 correct). Despite his poor understanding of mental state terms, the patient was perfectly able to infer other people's mental states in non-verbal tasks, including other people's intentions (he correctly inferred 23/28 intentions with Sarfati, Brunet, & Hardy-Bayle's, 2003, material), other people's knowledge gained from various physical interactions with objects (he scored 21/24 and 23/24 correct on tasks adapted from those used in developmental research, O'Neill, Astington, & Flavell, 1992; Pillow, 1993), as well as other people's beliefs (he scored 11/12 correct for both the reality-known and reality-unknown versions of the false belief task described earlier). This profile suggests that a rich lexico-semantic repertoire of terms denoting mental states is not necessary to reason about people's mental states.

Another important question regarding ToM knowledge relates to the organization in the mind and brain of ToM concepts compared to other types of concepts (e.g. other social concepts, such as those referring to personality traits, moral, and social conventions, or non-social concepts such as those related to the physical world). Several neuroimaging studies have found that some brain regions are more activated for social than non-social concepts suggesting that conceptual knowledge might be organized in the mind and brain along a social/non-social dimension. However, these studies offer a limited insight into our question of interest, for several reasons. First, these studies have led to different conclusions regarding the localization of social knowledge, some localizing it in the temporal poles (e.g. Ross & Olson, 2010; Simmons, Reddish, Bellgowan, & Martin, 2010; Zahn, Moll, Krueger, Huey, Garrido, & Grafman, 2007), whereas others localized it in the medial prefrontal cortex (Mason, Banfield, & Macrae, 2004; Mitchell, Heatherton, & Macrae, 2002; Mitchell, Bajani & Macrae, 2005). Secondly, ToM concepts have actually not been specifically (if at all) tested in these studies. The "social concepts" have mainly been contrasted with non-social concepts, without any investigation of how mental state concepts specifically are organized relative to other social or non-social concepts. Thus, how ToM-related concepts are organized in the mind and brain remains an open question. Finally, preferential activity observed in a brain region for a specific semantic category does not vouch for the *necessary* role of this brain region for the given semantic category.

The neuropsychological approach offers a way to overcome these limitations. First, neuropsychological studies offer a way to disentangle the different possibilities regarding the localization of social knowledge in the brain by assessing how social knowledge is specifically affected by a brain lesion in the putative areas. Zahn and colleagues (Zahn, Moll, Iyengar, Huey, Tierney, Krueger, et al., 2009) recently conducted such a study showing that a dysfunction of the right temporal pole was associated with a disproportionate impairment in processing social concepts compared to non-human animal function concepts. The authors concluded that the right temporal pole is necessary for representing conceptual social knowledge. As far as the left temporal pole is concerned, patient CM's good performance on ToM tasks despite a massive atrophy of the left temporal pole (Michel et al., 2011) suggests that at least ToM-related social concepts necessary to infer intentions, knowledge and beliefs would not be subtended by the left temporal pole (or at least not by the left temporal pole alone). Thus, if a repository for social knowledge was localized in the temporal pole as some authors proposed, it seems that it might be right lateralized or represented bilaterally.

Neuropsychological studies might also bring some answers to the other questions that have been left open by neuroimaging studies. The "social" category used in neuroimaging studies, as well as in Zahn et al. (2009) study, might be further divided in subcategories such as ToM-related concepts and various other kinds of social conceptual knowledge. Any selective deficit in patients would reflect that the dimension along which the affected and the preserved concepts differ could be a relevant principle for how knowledge is organized in the mind and brain, and that the brain area affected is necessary for the (most) impaired category of concepts. To the best of our knowledge, there is no such evidence in the literature, and the data collected on patient CM (Michel et al., 2011) were not informative as the patient was equally impaired for the various categories of concepts tested when these were presented verbally (i.e. there was no difference globally between social and non-social concepts, and no difference between mental states and personality trait concepts).

To summarize, the kind of ToM knowledge necessary to understand other people's minds and how this knowledge is organized in the mind and brain remain open questions for future research.

Low-level and high-level components to understand other people's minds

The different building blocks highlighted in the preceding sections show that using ToM to understand other people's minds requires the use of conceptual knowledge, the support of working memory and the recruitment of various inferential processes. Collectively, this shows that ToM is made of high-level cognitive components. It is thus not surprising that it takes time for children to become efficient at using their ToM and furthermore, that their ToM development occurs alongside important changes in their executive function and language abilities (Carlson & Moses, 2001; de Villiers & Pyers, 2002). This can also explain why even healthy adults do not necessarily automatically track other people's mental states (Apperly, 2006a) and why they often succumb to egocentric biases (Bernstein, Erdfelder, Meltzoff, Peria, & Loftus, 2011; Birch & Bloom, 2007; Epley et al., 2004; Keysar et al., 2000). However, this is not to suggest that we can only understand other people's minds by using our ToM or that using our ToM necessarily requires the full use of high-level cognitive representations and top-down processes. Indeed, in some circumstances taking into account someone else's visual experience, intention, emotion or even false belief can be facilitated by low-level processes which can give automatically and effortlessly useful information related to other people's mental states (see (e) in Figure 10.1). This useful information could be

acquired by bypassing the ToM components described earlier (see (a–d) in Figure 10.1) and/or could provide an input to the ToM components to support ToM reasoning.

One type of low-level process that can help us understand other people's mental states is that of visuo-motor priming sustained by the "mirror neuron system" (see Chapters 14 and 15). Through these visuo-motor priming mechanisms, observing someone else performing an action automatically activates a motor representation similar to the one we would activate ourselves if we were to do the action (e.g. Buccino, Binkofski, Fink, Fadiga, Gallese, Fogassi, et al., 2001). This mechanism might provide rich information about the other person's motor plan and could perhaps be used to predict what the other person will do. Similarly, there is evidence for emotional contagion effects whereby we automatically share other people's emotional states when we see others displaying these states behaviorally (e.g. Preston & de Waal, 2002; see also Chapters 12 and 13). In these cases, we could easily gain rich and useful information about other people's affective states. Finally, other low-level processes automatically draw our attention to what other people are looking at (Allison, McCarthy, & Puce, 2000; Driver, Davis, Ricciardelli, Kidd, Maxwell, & Baron-Cohen, S., 1999). This would provide useful information about their visual experience (Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010) and their preferences toward objects in the environment (Tipper, 2010).

The automaticity of these low-level processes means that some information about other people's mental states can be gained without accessing complex conceptual knowledge or monitoring the environment in a top-down manner, and thus bypassing some of the ToM components described earlier (see (b) and (d) in Figure 10.1). Furthermore, the automaticity of these low-level processes provides a way of enhancing the salience of the content of other people's mental states without having to inhibit one's own perspective, and thus bypassing another ToM component described earlier (see (a) in Figure 10.1). Hence, it might be that these low-level processes offer a way to understand some aspects of other people's minds and that they may be sufficient in some circumstances to interact with other people without the need to engage in complex and effortful computations. This may explain some of the intriguing recent findings showing that infants and even non-human animals are sensitive to other individuals' mental states (Emery & Clayton, 2009; Hare, Call, Agnetta, & Tomasello, 2000; Kovacs, Teglas, & Endress, 2010; Onishi & Baillargeon, 2005). Furthermore, if spared by their brain lesions, these low-level processes may give to patients with ToM impairments valuable support for their social interactions. However, such low-level processes alone do not provide the flexibility and scope to understand, reason and talk about the subtlety of people's mental states.

It is also possible that in some circumstances low-level processes interact with the high-level ToM components by providing useful input to ToM reasoning (Frith & Frith, 2012). They may even support the development of ToM components during childhood. If this is the case, impairments to the low-level processes may have cascading effects on the development and efficient use of ToM components. Whether, when and how low-level and high-level components interact is an open question for future research.

Conclusions

One of the current challenges in ToM research is to unravel the functional and neural architecture of the representations and processes which allow us to ascribe mental states. We hope to have illustrated that the study of the impaired and spared abilities of patients with acquired brain damage offers a valuable source of evidence to build such architecture and to specify the relations between the different building blocks with more general executive function and language processes.

We have discussed four types of building blocks that could form the basis of the ToM architecture, with each being independently represented in the brain (and mind) so that a lesion (or dysfunction) could selectively impair one of these building blocks while leaving the others spared. The first two building blocks refer to processes that appear executive in nature and that play a core role in the ToM inferential processes. One type of process allows us to resist interference from our own perspective enabling us to consider other people's discrepant mental states. The other type of process allows us to monitor the environment to detect and integrate relevant cues enabling us to give a specific content to other people's mental states. Further research is needed to characterize more precisely these processes and identify their neural substrates.

The third building block is a working memory support to represent and temporarily hold in our mind the content of other people's (as well as our own) mental states. Grammar (in the linguistic sense) does not seem to play a critical role in supporting the way mental states are represented in working memory. Thus, the question for future research is how we represent mental states, especially complex ones such as propositional attitudes that consist of embedded information.

The fourth building block consists of a system storing ToM knowledge in long-term memory. Preliminary findings indicate that the social/non-social distinction may be a relevant dimension along which our conceptual knowledge is organized in the mind and brain, but more studies are needed to specifically investigate how ToM knowledge is organized relative to other types of social and non-social knowledge. Even if we progress in understanding the topography of conceptual knowledge, it will be important to determine the minimum knowledge necessary to infer other people's mental states. The only related evidence so far is that we don't need to be able to use or understand a rich and extensive repertoire of terms denoting mental state terms to reason about mental states.

These four building blocks support the ToM reasoning that accompanies explicit mentalizing. However, we highlighted that we also have other, more low-level processes by which valuable information about other people's minds may be gained. Information gained by these processes may not be construed as mental states, but may help in interacting efficiently with other people without the need to recruit effortful processes and complex representations. Whether, when and how low-level and high-level components interact are important questions for future research. Similarly to many other open questions highlighted in this chapter, neuropsychological findings may continue to be one valuable source of evidence to find the answers to these questions.

References

- Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: falseness, beliefs, or attention. *Journal of Cognitive Neuroscience* 21(6): 1179–92.
- Allison, T., McCarthy, G., & Puce, A. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences* 4(7): 267–78.
- Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science* 17(10): 841–4.
- Apperly, I. A., Samson, D., Carroll, N., Hussain, S., & Humphreys, G. W. (2006). Intact first- and second-order false belief reasoning in a patient with severely impaired grammar. *Social Neuroscience* 1(3–4): 334–48.
- Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience* 16: 1773–84.
- Apperly, I. A., Samson, D., Chiavarino, C., Bickerton, W.-L., & Humphreys, G. W. (2007). Testing the domain-specificity of a theory of mind deficit in brain-injured patients: evidence for consistent

- performance on non-verbal, "reality-unknown" false belief and false photograph tasks. *Cognition* 103(2): 300–21.
- Bailey, P. E., & Henry, J. D., 2008. Growing less empathic with age: disinhibition of the self-perspective. *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, 63(4): 219–26.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition* 21(1): 37–46.
- Bernstein, D. M., Erdfelder, E., Meltzoff, A. N., Peria, W., & Loftus, G. R. (2011). Hindsight bias from 3 to 95 years of age. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(2): 378–91.
- Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science* 18(5): 382–7.
- Bodden, M. E., Dodel, R., & Kalbe, E. (2010). Theory of mind in Parkinson's disease and related basal ganglia disorders: a systematic review. *Movement Disorders* 25(1): 13–27.
- Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Gallese, V., Fogassi, L., et al. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience* 13(2): 400–4.
- Burgess, P. W., Dumontheil, I., & Gilbert, S. J. (2007). The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences* 11(7): 290–8.
- Call, J., & Tomasello, M. (1999). A non-verbal false belief task: the performance of children and great apes. *Child Development* 70(2): 381–95.
- Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development* 72(4): 1032–53.
- Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition* 6(5): 509–40.
- Emery, N. J., & Clayton, N. S. (2009). Comparative social cognition. *Annual Review of Psychology* 60: 87–113.
- Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology* 87(3): pp.327–39.
- Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, 63: 287–313.
- Gopnik, A. (2003). The theory theory as an alternative to the innateness hypothesis. In L. Antony & N. Hornstein (Eds), *Chomsky and His Critics* (pp. 238–254). Oxford: Blackwell Publishing.
- Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind and Language* 7(1–2): 145–71.
- Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language* 1(2): 158–71.
- Happé, F., Brownell, H., & Winner, E. (1999). Acquired "theory of mind" impairments following stroke. *Cognition* 70(3): 211–40.
- Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour* 59(4): 771–85.
- Houthuys, S. (2010). The functional and neural basis of belief and desire reasoning. Unpublished PhD thesis, University of Birmingham.
- Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science* 11(1): 32–8.
- Kovacs, A. M., Teglas, E., & Endress, A. D. (2010). The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science* 330(6012): 1830–4.
- Lough, S., & Hodges, J. R. (2002). Measuring and modifying abnormal social cognition in frontal variant frontotemporal dementia. *Journal of Psychosomatic Research* 53(2): 639–46.
- Mason, M. F., Banfield, J. F., & Macrae, C. N. (2004). Thinking about actions: the neural substrates of person knowledge. *Cerebral Cortex* 14(2): 209–14.
- McDonald, S., & Bibby, H. (2005). Theory of mind after traumatic brain injury. *Neuropsychologia*, 43(1): 99–114.

- McKinnon, M. C., & Moscovitch, M. (2007). Domain-general contributions to social reasoning: theory of mind and deontic reasoning re-explored. *Cognition*, 102(2): 179–218.
- Michel, C., Dricot, L., Lhommel, L., Grandin, C., Ivanoiu, A., Pillon, A. et al. (2011). On the role of the left anterior temporal lobe in social cognition: Neuropsychological evidence from a patient with semantic dementia. Poster presented at the *Social Brain Symposium*, Bruxelles, Belgium, 14 November 2011.
- Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *NeuroImage* 28(4): 757–62.
- Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences of the USA*, 99(23), pp.15238–43.
- van der Meer, L., Groenewold, N., Nolen, W., Pijnenborg, M., & Aleman, A. (2011). Inhibit your self and understand the other: Neural basis of distinct processes underlying theory of mind. *NeuroImage* 56(4): 2364–74.
- O'Neill, D. K., Astington, J. W., & Flavell, J. H. (1992). Young children's understanding of the role that sensory experiences play in knowledge acquisition. *Child Development* 63(2), 474–90.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science* 308(5719): 255–8.
- Peterson, C. C., & Siegal, M. (2000). Insights into theory of mind from deafness and autism. *Mind and Language* 15(1): 123–45.
- Pillow, B. H. (1993). Preschool children's understanding of the relationship between modality of perceptual access and knowledge of perceptual properties. *British Journal of Developmental Psychology* 11(4): 371–89.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 4: 515–26.
- Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences* 25(1): 1–20.
- Ross, L. A., & Olson, I. R. (2010). Social cognition and the anterior temporal lobes. *NeuroImage* 49(4): 3452–62.
- Samson, D. (2009). Reading other people's mind: insights from neuropsychology. *Journal of Neuropsychology* 3(1): 3–16.
- Samson, D., Apperly, I. A., & Humphreys, G. W. (2007). Error analyses reveal contrasting deficits in “theory of mind”: neuropsychological evidence from a 3-option false belief task. *Neuropsychologia* 45(11): 2561–69.
- Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology. Human Perception and Performance*, 36(5): 1255–66.
- Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nature Neuroscience* 7(5): 499–500.
- Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: a case of a selective deficit in inhibiting self-perspective. *Brain*, 128: 1102–11.
- Sarfati, Y., Brunet, E., & Hardy-Baylé, M. C. (2003). *Comic-strip Task: Attribution of Intentions to Others*. Service de Psychiatrie Adulte, Hôpital de Versailles, Le Chesnay, France.
- Siegal, M., Carrington, J., & Radel, M. (1996). Theory of mind and pragmatic understanding following right hemisphere damage. *Brain and Language* 53(1): 40–50.
- Simmons, W. K., Reddish, M., Bellgowan, P. S. F., & Martin, A. (2010). The selectivity and functional connectivity of the anterior temporal lobes. *Cerebral Cortex* 20(4): 813–25.
- Smith, M., Apperly, I. A., & White, V. (2003). False belief reasoning and the acquisition of relative clause sentences. *Child Development* 74(6): 1709–19.
- Sodian, B., & Thoermer, C. (2008). Precursors to a theory of mind in infancy: perspectives for research on autism. *Quarterly Journal of Experimental Psychology* 61(1): 27–39.

- Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience* 10(5): 640–56.
- Stone, V. E., Baron-Cohen, S., Calder, A. J., Young, A. W., & Keane, J. (2003). Acquired theory of mind impairments in individuals with bilateral amygdala lesions. *Neuropsychologia* 41(2): 209–20.
- Stuss, D. T., Gallup, G. G., & Alexander, M. P. (2001). The frontal lobes are necessary for “theory of mind.” *Brain* 124: 279–86.
- Surian, L., & Siegal, M. (2001). Sources of performance on theory of mind tasks in right hemisphere-damaged patients. *Brain and Language* 78(2): 224–32.
- Tipper, S. P. (2010). From observation to action simulation: the role of attention, eye-gaze, emotion, and body state. *Quarterly Journal of Experimental Psychology* 63(11): 2081–105.
- Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping* 30(3): 829–58.
- Varley, R., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and “theory of mind” in an agrammatic aphasic patient. *Current Biology* 10(12): 723–6.
- Varley, R., Siegal, M., & Want, S. C. (2001). Severe impairment in grammar does not preclude theory of mind. *Neurocase* 7(6): 489–93.
- Villiers, J. G. D., & Pyers, J. E. (2002). Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development* 17: 1037–60.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition* 13(1): 103–28.
- Zahn, R., Moll, J., Iyengar, V., Huey, E. D., Tierney, M., Krueger, F., et al. (2009). Social conceptual impairments in frontotemporal lobar degeneration with right anterior temporal hypometabolism. *Brain* 132(3): 604–16.
- Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences of the USA* 104(15): 6430–35.

Understanding emotional and cognitive empathy: A neuropsychological perspective

Anat Perry and Simone Shamay-Tsoory

Awful disasters descended upon FJ. She was happily married to MJ, and they had three lovely children, until one day MJ said he decided to leave her. But that was only the beginning—for the next few months he started verbally abusing her and her children (which he never did before), left the house, abandoned all his responsibilities to others—taking the children from school, answering phone calls, cutting back on work, etc. FJ refused to accept that her marriage had suddenly collapsed, and that it was beyond mending. Whatever FJ tried, she felt she just couldn't get through to MJ. For friends around them, this was another typical divorce story, but FJ kept saying she felt MJ has gradually become a completely different person—this was not the MJ she fell in love with, married and raised a family with. They went through a terrible divorce, at times involving the police and social workers in the community. A few months following the divorce, MJ was hospitalized with a chronic headache and a CT scan found a large tumor in his frontal lobe, which had probably been there and growing for the last couple of years. In a way, FJ had been right; this wasn't the MJ she married.

Probably every first year Psychology or Neuroscience student is familiar with the famous case report of Phineas Gage, which presented one of the first descriptions of aberrant social conduct following brain damage (Harlow, 1868). Neuroscience research since Harlow's time has shown that impaired empathic abilities among people with different brain lesions, tumors and pathologies may account in part for the social and behavioral disturbances often observed among such patients. These seminal studies have a two-way effect—they further our knowledge of brain networks that enable empathy, and they may contribute to future diagnosis and therapeutic development of different social deficits and pathologies. Importantly, progress has been made in the past decade, both in the psychological terminology and through neuroimaging studies, in differentiating between sub-components of empathy, i.e. emotional and cognitive empathy, and further differentiating cognitive empathy to its affective and cognitive components. This chapter will provide a theoretical framework for understanding of emotional and cognitive empathy. Combining the findings from lesion studies, electrophysiology and imaging data of healthy and patient groups, a neuroanatomical model for empathy and its relationship with theory of mind will be proposed.

How do we normally empathize with others? A psychological background

Empathy is a broad concept, which generally denotes our ability to identify with or to feel what the other is feeling. In order to empathize with the other, we need to first understand *what* that person

is going through, feeling or thinking. Two main psychological approaches have dominated the discussion of how we understand others' minds—simulation-theory and theory-theory (Carruthers, 1996; Davies & Stone, 1995). Simulation-theory suggests that we use our own mental mechanisms to estimate and predict the mental processes of others. According to this theory, we automatically create within ourselves a simulation of the others' motor acts, and in accordance with these acts, we feel the desires, preferences, and beliefs of the sort we assume the other to have. Instead of acting on that decision, it is taken "off-line," and used to predict the intention and emotions of the other (Gallese & Goldman, 1998; Goldman, 1989; Gordon, 1986; Heal, 1986; Kosslyn, 1978; Stich, 1992).

The second theory, theory-theory, posits that children, as they grow up, use the same cognitive mechanisms that adults use in science, that is they develop theories. These theories help them investigate the world around them, infer from one situation to the next, and predict others' behaviors and emotions. In other words, understanding others and empathizing with them arise from theoretical reasoning involving known causal laws, or knowledge of how people in general, or people with the other's specific characteristics, are likely to think or feel (Churchland, 1988; Fodor, 1987; Gopnik, 1993; Lewis, 1972; Wellman, 1990). This ability, of attributing mental states such as beliefs, intents, desires, pretending, knowledge, etc., to oneself and others, and understanding that others have beliefs, desires and intentions that may be different from one's own has also been labeled theory of mind (ToM; Frith & Frith, 1999; Premack & Woodruff, 1978).

Developmental studies indicate that emotional contagion (e.g. contagious crying) is observed very early in young babies (e.g. Simner, 1971), while cognitive perspective taking abilities, ascribing agency, understanding intentions, and joint attention skills are acquired during later cognitive development (e.g. Carpenter, Nagell, and Tomasello, 1998; Johnson, 2003). More advanced ToM skills usually appear only by 3 or 4 years of age. These include understanding "false-belief", i.e. that the other may have a belief that you know is false (e.g. (Flavell, Everett, Croft, & Flavell, 1981; Flavell, Flavell, & Green, 1983; Wimmer & Perner, 1983). In a typical false-belief task (also known as the "Sally-Anne task"), the child is asked where a person (e.g. Sally) will look for a toy that she left in a certain place and that was moved by another person (e.g. Anne) when Sally was out of the room. Understanding that Sally will not know the new location of the toy relies on the ability to distinguish between Sally's false belief and reality (Wimmer & Perner, 1983).

Current evolutionary evidence also supports the existence of several systems mediating empathy. De Waal (2008) suggests that the phylogenetically earliest system is the emotional contagion system, in which one is affected by another's emotional or arousal state. On the other hand, the cognitive empathic perspective-taking system is a more advanced system and involves higher cognitive functions including mental state attribution. Indeed, emotional contagion has been reported in rodents (Langford, Cragger, Shehzad, Smith, Sotocinal, Levenstadt, et al., 2006), while only the closest living relatives of humans, the chimpanzees, possess rudimentary traits of cognitive aspects of empathy such as theory of mind (Call & Tomasello, 2008).

Indeed, empathy is a broad concept that refers to the cognitive, as well as the emotional reactions of one individual to the observed experiences of another. Neuroimaging, lesion and behavioral studies with humans and animals have been increasingly capable of characterizing the neural basis of empathy, thus providing new insights into the question of how we understand others' minds. Recent evidence supports a model of two separate brain systems for empathy: an emotional system and a cognitive system. The capacity to experience affective reactions to the observed experiences of others or share a "fellow feeling" has been described as "emotional empathy". Emotional empathy may involve several related underlying processes, including, among others, emotional contagion, emotion recognition, and shared pain. This might be where a kind of "simulation theory" takes.

On the other hand, the term “cognitive empathy” describes empathy as a cognitive “role-taking” ability, or the capacity to engage in the cognitive process of adopting another’s psychological point of view (Frith & Singer 2008). This ability may involve a “theory-theory” kind of reasoning, i.e. making inference regarding the other’s affective and cognitive mental states (Shamay-Tsoory, Aharon-Peretz & Perry, 2009).

This suggests that, while the two systems may work together, they may be behaviorally, developmentally and neuroanatomically dissociable. Here, we will focus on these two systems from a neuropsychological point of view—reviewing the current research from imaging, lesion studies, and different pathologies related to empathic abilities.

Emotional empathy

According to Preston & de Waal’s (2002) perception-action hypothesis, perception of a behavior in another automatically activates one’s own representations for the behavior, and output from this shared representation automatically proceeds to motor areas of the brain where responses are prepared and executed. This state-matching reaction has been related to the simulation theory. The most prominent evidence for the biological feasibility of the simulation theory is the seminal discovery of mirror neurons in the macaque monkey, a particular class of visuo-motor neurons that discharge both when the monkey does a particular goal-directed action and when it observes another individual (monkey or human) doing a similar action (di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). Such neurons have been discovered primarily in areas F5 of the monkey’s pre-motor cortex (roughly corresponding in humans to Brodmann (BA) areas 44 and parts of BA 45), in the rostral part of the inferior parietal lobule (IPL, particularly in area PFG) and in the anterior intraparietal sulci (AIP) (Fogassi, Ferrari, Gesierich, Rozzi, Chersi, & Rizzolatti, 2005; Rozzi, Ferrari, Bonini, Rizzolatti, & Fogassi, 2008). Given its observation-execution properties, it was suggested that the mirror neuron system (MNS) is particularly well suited to provide the appropriate mechanism for imitation, emotional contagion, and by extension, enabling empathy.

Neuroimaging studies in humans found different brain regions that similarly to the MNS in the monkey, are activated on the one hand by motor performance and, on the other hand, by seeing similar movements made by others. Such activity was found primarily in the rostral part of the IPL, the lower part of the precentral gyrus and the posterior part of the inferior frontal gyrus (IFG) (Buccino, Binkofski, Fink, Fadiga, Fogassi, Gallese, et al., 2001; Decety, Chaminade, Grezes, & Meltzoff, 2002; Grafton, Arbib, Fadiga, & Rizzolatti, 1996; Grezes, Armony, Rowe, & Passingham, 2003; Grezes, Costes, & Decety, 1998; Iacoboni et al., 1999; Rizzolatti et al., 1996). There is consistent and strong evidence for the involvement of the IFG in emotional contagion and emotion recognition. Chakrabarti, Bullmore, & Baron-Cohen (2006) found a positive correlation between a validated measure of empathy (the empathy quotient, EQ) and IFG activation, when viewing video clips depicting basic emotions. Indeed, it has been suggested that overt facial mimicry (as measured by an electro-myograph or through observation) is related to emotional contagion and emotion understanding (Niedenthal, 2007). The existence of mirror neurons related to emotional facial expressions in the human IFG suggests that the human MNS may be used to convert observed facial expressions into a pattern of neural activity that would be suitable for producing similar facial expressions and provide the neural basis for emotional contagion (Keysers & Gazzola, 2009). Jabbi and colleagues (Jabbi, Swart, & Keysers, 2007) have reported that observing positive and disgust facial expressions activated parts of the IFG and that participants’ empathy scores were predictive of their IFG activation while witnessing facial expressions. Additionally, two

neuroimaging studies, one which involved emotion recognition (Schulte-Ruther, Markowitsch, Fink, & Piefke, 2007) and one that involved empathizing with people suffering serious threat or harm (Nummenmaa, Hirvonen, Parkkola, & Hietanen, 2008) have further emphasized the specific role of the IFG in emotional empathy. Cortical lesions involving the IFG, particularly in BA 44, are associated with impaired emotional contagion and deficits in emotion recognition, suggesting that the IFG not only participates in tasks that involve emotional empathy, but is also **necessary** for emotional empathy. Shamay-Tsoory et al. (2009) studied eight patients with IFG damage, and compared their results with those of ventro-medial (VM) lesion patients, posterior lesion patients and healthy controls. The IFG patients differed from the other groups in the emotional empathy subscales of the Interpersonal Reactivity Index (IRI, Davis, 1983), an empathy questionnaire that differentiates between cognitive and emotional components of empathy. In addition, these patients were significantly worse in an emotional recognition task, but did not differ from healthy individuals in the ToM tasks. The authors show a double dissociation between IFG patients and VM patients, who showed significantly different results in the cognitive subscale and ToM task (Shamay-Tsoory et al., 2009).

A different neural manifestation that has been tentatively associated with the MNS in humans is a modulation of electroencephalographic (EEG) oscillations in the 8–13 frequency range, coined mu rhythms (Pineda, 2005). Mu rhythms are desynchronized and their power attenuated when engaging in motor activity (Gastaut, 1952) and, crucially, also while observing actions executed by someone else (Cochin, Barthelemy, Lejeune, Roux, & Martineau, 1998; Cochin, Barthelemy, Roux, & Martineau, 1999; Cohen-Seat, Gastaut, Faure, & Heuyer, 1954; Gastaut & Bert, 1954; Muthukumaraswamy, Johnson, & McNair, 2004). The visual-motor coupling suggested by this pattern led several investigators to suggest that it reflects a “resonance system”, simulating others’ motor actions (see Pineda, 2005, for a review). In the last few years, several studies linked mu suppression to higher social behaviors, such as understanding others social interactions (Oberman, Pineda, & Ramachandran, 2007), intentions (Perry, Troje, & Bentin, 2010) and empathy (Cheng, Lee, Yang, Lin, Hung, & Decety, 2008a; Cheng, Yang, Lin, Lee, & Decety, 2008b; Perry et al., 2010).

However, not only motor regions exhibit mirror-like mechanisms. Tactile processing mechanisms have also been shown to have mirror properties. The second and third somatosensory cortices can be vicariously recruited by the sight of other people being touched, performing actions or experiencing somatic pain (see Keysers, Kaas, & Gazzola, 2010, for a review). Smelling a foul odor engages the same sub-regions of the anterior insula as does watching another person express smell-induced disgust (Wicker, Keysers, Plailly, Royet, Gallese, & Rizzolatti, 2003). Shared pain appears to involve regions related to the first hand experience of pain, such as parts of the pain matrix. Specifically, a network including the anterior cingulate cortex (ACC) and the insula was reported to respond to both felt and observed pain (Decety, Yang, & Cheng, 2010). Activation in the ACC and insula has been also found to correlate with the participant’s judgments of the subjective severity of pain experienced by others on the basis of the other’s facial pain expression (Saarela, Markowitsch, Fink, & Piefke, 2007). This indicates that empathizing with people in pain is associated with hemodynamic activity in the brain that is similar to the activity that occurs when people feel pain themselves.

Finally, emotional empathy is also crucially dependent to the limbic system, and particularly on the amygdala. This brain structure, which is constructed of a complex collection of nuclei, is known for its substantial role in both experiencing and recognizing emotions, especially fear, and in social behavior and reward learning in general (see Adolphs, 2010, for a review). Adolphs and colleagues (Adolphs, Tranel, Hamann, Young, Calder, Phelps, et al., 1999) examined a cohort of nine individuals with rare bilateral amygdala damage. Compared with controls, the subjects as a

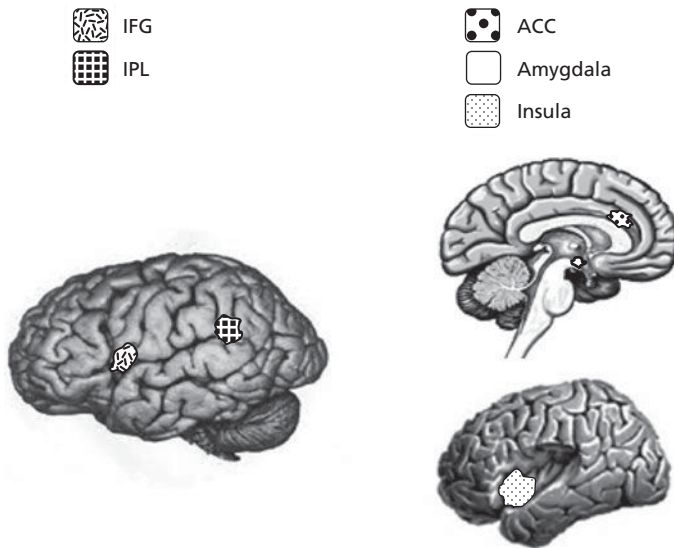


Figure 11.1 Brain regions associated with emotional empathy. See also Plate 3.

group were significantly impaired in recognizing fear, although individual performances ranged from severely impaired to essentially normal. An extensive line of research has been done on a specific patient, SM, with bilateral amygdale damage, who does not seem to recognize fear in others or to experience the feeling of fear or danger in herself (Adolphs, Tranel, Damasio, & Damasio, 1994). Impaired emotion recognition has also been linked to abnormal amygdala activation in psychiatric illnesses in which the behavioral relevance of social stimuli is abnormally evaluated, ranging from phobias (e.g. Larson, Schaefer, Siegle, Jackson, Anderle, & Davidson, 2006) to depression (e.g. Pezawas, 2005) schizophrenia (e.g. Das, Kemp, Flynn, Harris, Liddell, Whitford et al., 2007; Gur Calkins, Gur, Horan, Nuechterlein, & Seidman, 2007) and autism (e.g. Spezio, Adolphs, Hurley, & Piven, 2007). Interestingly, a similar impairment has also been found in first-degree relatives of people with autism and may constitute part of an “endophenotype” for impaired amygdala function in autism (Dalton Nacewicz, Alexander, & Davidson, 2007; Adolphs, Spezio, Parlier, & Piven, 2008).

To conclude, the core process enabling emotional empathy appears to be the generation of corresponding (to the target) emotional response (e.g. through the insula in shared pain, the amygdala in fear), and the corresponding motor representation related to the emotion (e.g. IFG, mu rhythms). The neural networks that participate in this system are detailed in Figure 11.1. Although this system appears to be bottom-up, it seems that top-down processes can modulate this automatic system and perhaps aspects of higher order cognitive process, as well as cognitive empathy, interact with emotional empathy in such cases (e.g. Lamm, Meltzoff, & Decety, 2010; Perry, Bentin, Ben-Ami Bartal, Lamm, & Decety, 2010).

Cognitive empathy

Some complex forms of empathy involve the ability to create a theory about the other’s mental state and cognitively take the perspective of others. This process of understanding another person’s perspective, termed “cognitive empathy” is very much related to having ToM abilities. While one

dimension of ToM relates to others' beliefs and desires, another dimension concerns the emotional and social meaning of others' intentions (Brothers & Ring, 1992). Brothers & Ring (1992) referred to these dimensions as "cold" and "hot" aspects of ToM, and suggested that both forms of cognition contribute to understanding others' actions. The distinction between "cold" aspects of mental representations (cognitive ToM) as opposed to "hot" aspects of mental representations (affective ToM), has been further extended and examined in lesion studies (e.g. Shamay-Tsoory Tomer, Berger, Goldsher, & Aharon-Peretz, 2005; Shamay-Tsoory & Aharon-Peretz, 2007) and functional imaging studies (Hynes, Baird, & Grafton, 2006; Völlm, Taylor, Richardson, Corcoran, Stirling, McKie, S., et al., 2006). These have differentiated between cognitive and affective ToM, referring to reasoning about beliefs vs. reasoning about emotions, respectively. If we take the example of "false belief," cognitive false belief requires understanding what someone thinks about what someone else thinks (belief about belief), while affective false belief refers to understanding what someone thinks about what someone else feels (belief about emotions). Consistent with the possibility that ToM skills comprise several distinct processes that meet different cognitive demands, recent studies have identified a set of brain regions involved in ToM: the medial prefrontal cortex (mPFC), the superior temporal sulcus (STS), the temporoparietal junction (TPJ) and the temporal poles (TP) (Frith & Singer, 2008; Van Overwalle & Baetens, 2009). A recent review of imaging studies of ToM, (Carrington & Bailey, 2009) found that 93% of the 40 studies reviewed report activation in the mPFC. The TPJ region was active in 58% of the studies reviewed and the STS (including the IPL) in 50% of the studies. Based on a separate meta-analysis, Van Overwalle & Baetens (2009) proposed that the TPJ is mainly responsible for transient mental inferences about other people (e.g. their goals, desires, and beliefs), while the mPFC subserves the attribution of more enduring traits and qualities about the self and other people. The ventromedial prefrontal cortex (vmPFC) role in self-reflection has been lately shown in an fMRI study by Mitchell and colleagues (Mitchell, 2009), placing it as a key region necessary for evaluating the similarities and differences distinguishing the mental states of oneself from others. It is possible that situations that involve affective ToM entail more self-reflection as compared with situations involving cognitive ToM, which are more detached. Therefore the vmPFC, which is highly connected to the amygdala, appears to be particularly necessary for affective mentalizing, as opposed to neutral or cognitive forms of mentalizing, along with mentalizing about others who are different from oneself, which have been linked to the modulation of dorsal regions of the mPFC (dmPFC; e.g. Mitchell, Macrae and Banaji, 2006). In addition, Kalbe and colleagues (Kalbe, Schlegel, Sack, Nowak, Dafotakis, Bangard, et al., 2010) have recently reported impaired cognitive ToM, following 1Hz repetitive Transcranial Magnetic Stimulation (TMS) which interfered with cortical activity of the dorsolateral PFC (dlPFC).

The frontal lobes, especially regions of the PFC, have been associated with executive aspects of cognition, social and moral behavior ever since description of frontal lobe syndromes-related changes in personality, social behavior, and emotional regulation emerged in the nineteenth century (Eslinger, Flaherty-Craig, & Benton, 2004). There is ample evidence that the orbitofrontal cortex (OFC) mediates affective information, emotional stimuli, and social behavior. Lesions in the OFC result, among other things, in impaired empathy (Eslinger, 1998; Shamay-Tsoory, Tomer, Goldsher, Berger, & Aharon-Peretz, 2004) and deficits in complex ToM abilities (Stone Baron-Cohen, & Knight, 1998). Shamay-Tsoory, Tibi-Elhanani, & Aharon-Peretz (2006) compared the performance of patients with lesions localized either in the vmPFC (part of the OFC), dorsolateral, TPJ, or superior parietal to healthy controls, with a battery of naturalistic affective and cognitive ToM stories (about false beliefs, false attribution, irony and lies). Compared with controls, patients with VM damage were impaired at providing appropriate mental state explanations for the affective ToM stories (see also Shamay-Tsoory et al., 2009). In addition, these patients'

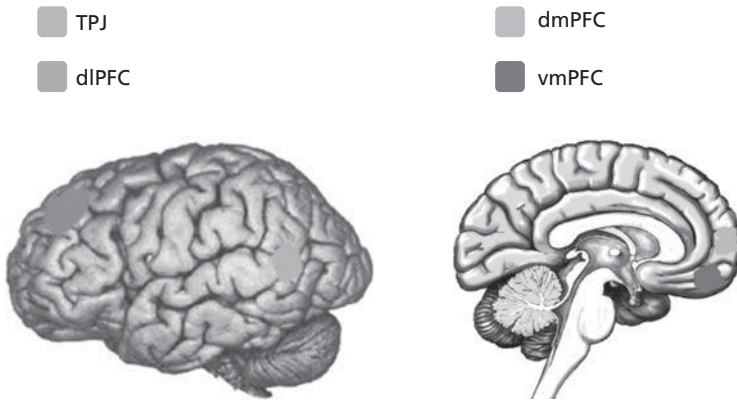


Figure 11.2 Brain regions associated with cognitive empathy. See also Plate 4.

performance in the affective ToM was significantly impaired as compared with their performance in cognitive ToM stories. Furthermore, ratings of levels of emotionality of each story suggested that levels of affective load correlated with number of errors in the stories in the VM group, indicating that the more the emotional load involved in the story the greater the difficulty posed for the subjects in this group (see Figure 11.2).

To conclude, it appears that cognitive empathy involves higher order cognitive functions that involve self–other differentiation, and cognitive and affective ToM. Self-other distinction and affective ToM involve a network in which the OFC (specifically vmPFC), and the TPJ to some extent, are core regions. The dIPFC and dmPFC also seem to be involved in cognitive ToM, although more studies are needed to further dissociate brain regions necessary for each sub-component of ToM.

Neuropsychiatric deficits in empathy

There has been a marked increase in empathy and ToM studies in psychiatric populations that examine the pattern of brain activity in response to social perception tasks among patients. The failure of most studies to find clear activation or deactivation in specific brain regions, indicates the difficulties in recruiting a homogenous patient group, adjusting the behavioral measures and finding the appropriate control groups. Nonetheless, studying empathy in psychiatric populations may have important contributions to both basic research and to diagnosis and therapy. Some of these contributions are discussed below.

Autism

Autistic spectrum disorders (ASD) include a wide range of deficits in social cognitive skills, including deficits in ToM abilities, i.e. in understanding thoughts and intentions of others, and differentiating between one’s own knowledge and the knowledge of the other (e.g. Baron Cohen, 2000). In an fMRI study, Lombardo and colleagues (Lombardo, Chakrabarti, Bullmore, Sadek, Pasco, Wheelwright, et al., 2010) have demonstrated that while typical individuals recruit the ACC and the vmPFC in response to self and other-referential processing respectively, in autism, vmPFC responded equally to self and other. Furthermore, the magnitude of neural self-other distinction in ventromedial prefrontal cortex was strongly related to the magnitude of early childhood social impairments in autism. Individuals whose vmPFC made little to no distinction between

mentalizing about self and other were the most socially impaired in early childhood. In a different fMRI study (Schulte-Ruther, et al., 2011), individuals with autism and healthy control subjects were asked to identify the emotional state observed in a facial stimulus (other-task) or to evaluate their own emotional response (self-task). Activations in these tasks were found in vmPFC only in control subjects, while more dorsal activation was found in ASD subjects. During the self-task, ASD subjects activated an additional network of frontal and inferior temporal areas. The authors concluded that subjects with ASD may use an atypical cognitive strategy to gain access to their own emotional state in response to other people's emotions. The differences between the results of these two studies may be due to the different tasks used, or to a different subgroup of ASD individuals. However, both studies strengthen the notion that this complex disorder is correlated with abnormalities in brain regions that enable our understanding of others, and the differentiation between self and others.

Mirror neuron system dysfunction and autism

To examine MNS abnormalities in individuals with ASD, Dapretto and colleagues studied high functioning children with ASD and matched controls, in an fMRI experiment involving imitation and observation of emotional expressions. Although there were no behavioral differences between the groups, children with ASD showed no mirror activity in the IFG. Moreover, activity in this area was inversely related to symptom severity in the social domain (Dapretto, Davies, Pfeifer, Scott, Sigman, Bookheimer, et al., 2006; see also Uddin, Davies, Scott, Zaidel, Bookheimer, Iacoboni, et al., 2008; Schulte-Ruther, Greimel, Markowitsch, Kamp-Becker, Remschmidt, Fink, et al., 2010 for similar results).

In line with this, studies of EEG mu suppression in individuals with ASD show normal mu suppression while self-performing hand movements, but no suppression when passively viewing someone else performing the same movements (Martineau, Schmitz, Assaiante, Blanc, & Barthelemy, 2004; Oberman, Hubbard, McCleery, Altschuler, Ramachandran, & Pineda 2005; Oberman, Ramachandran, & Pineda, 2008; but see Raymaekers, Wiersema, & Roeyers, 2009). Oberman and colleagues (2008) found a positive correlation between the amount of mu suppression and the putative ability of the observer to identify with the agent moving on the screen: Both ASD and typically developed individuals showed greater suppression to familiar hands (of family members) compared with those of strangers. These studies strengthen the notion that ASD patients may have dysfunction in mirror neuron systems, however, further studies are needed to validate this claim and especially to understand what is the cause (MNS dysfunction, Autism) and what is the effect.

Schizophrenia

Schizophrenia constitutes a complex mental disorder, which includes what is typically referred to as "positive symptoms" (delusions, hallucinations, disorganized speech, disorganized or catatonic behavior), and/or "negative symptoms" (affective flattening, alogia, or anhedonia). Patients suffering from schizophrenia show impaired emotional and social behavior, such as misinterpretation of social situations and lack of ToM skills. Differentiating between cognitive and affective ToM, Shamay-Tsoory and colleagues (2007) showed that patients with schizophrenia made significantly more errors in the affective conditions, compared with controls. Furthermore, correlation analysis indicated that impaired affective ToM in these patients correlated with their level of negative symptoms. These results indicate that individuals with high level of negative symptoms of schizophrenia may demonstrate selective impairment in their ability to attribute affective mental states.

Following this study, the performance of 24 patients with schizophrenia was compared with the responses of patients with localized lesions in the VM or dorsolateral PFC, patients with non-frontal lesions, and healthy control subjects. Patients with schizophrenia and those with VM lesions made similar errors on “affective ToM” tasks showed normal results in cognitive ToM conditions. The authors concluded that the pattern of mentalizing impairments in schizophrenia resembled those seen in patients with lesions of the frontal lobe, particularly with VM damage, providing support for the notion of a disturbance of the fronto-limbic circuits in schizophrenia (Shamay-Tsoory, Aharon-Peretz, & Levkovitz, 2007). Strengthening this notion, an fMRI study using a modified emotional-Stroop task, showed that patients with schizophrenia were significantly less efficient than the healthy controls during the emotional incongruent Stroop trials, and when emotionally incongruent trials were compared with congruent trials, relative deactivations of the subgenual cingulate gyrus and the vmPFC observed in the healthy controls were not found in the patient group. Importantly, activities of these regions inversely correlated with emotional interference to performance efficiency and response accuracy respectively in the patient group (Park, Park, Chun, Kim, & Kim, 2008). Recently, an important study showed that ToM skills are related to gray matter volume (GMV) in the vmPFC in Schizophrenia (Hooker, Bruce, Lincoln, Fisher, & Vinogradov, 2011). The authors used voxel-based morphometry and a multi-method behavioral assessment of ToM processing, including performance-based, self-report, and interview-rated ToM assessments, to investigate whether ToM skills were related to vmPFC GMV. As expected, compared with healthy participants, schizophrenia participants had worse ToM performance and lower self-reported ToM processing in daily life. Importantly, schizophrenia participants had less vmPFC GMV than healthy participants. Moreover, among schizophrenia participants, all three measures of ToM processing were associated with vmPFC GMV, such that worse ToM skills were related to less VMPFC GMV. This association remained strong for self-reported and interview-rated ToM skills, even when controlling for the influence of global cognition. Together, these studies suggest a strong correlation between abnormalities in vmPFC, and cognitive empathy skills in schizophrenia patients.

MNS dysfunction and schizophrenia

The discovery of the MNS had led to much excitement, and was seen as an opportunity to further our understanding of different disorders of social cognition and empathy. It has been suggested in the last few years that both the positive and the negative symptoms of schizophrenia may be due to a dysfunctional MNS (e.g. Arbib & Mundhenk, 2005; Buccino & Amore, 2008; Burns, 2006), although only little support has been found for these hypotheses (Enticott, Hoy, Herring, Johnston, Daskalakis, & Fitzgerald, 2008; Singh, Pineda, & Cadenhead, 2011) and more research is needed in order to understand the relationship between these brain networks and the complex disorder of schizophrenia.

Psychopathy

Patients with OFC lesions described above have sometimes been described as expressing “acquired sociopathy” (Blair & Cipolotti, 2000; Tranel, Bechara, & Denburg, 2002), a term denoting aberrant behavior, high levels of aggression, and a callous disregard for others following OFC lesions. Similarly, criminal offenders with psychopathic tendencies show impaired emotional and social behavior, such as lack of emotional responsiveness to others and deficient empathy. To assess the emotional and cognitive aspects of ToM in people with psychopathic tendencies, Shamay-Tsoory, Harari, Aharon-Peretz, & Levkovitz (2010) used a task that examines affective vs. cognitive ToM processing in separate conditions. ToM abilities of criminal offender diagnosed with antisocial

personality disorder with high psychopathy features was compared with that of participants with localized lesions in the OFC or dorsolateral participants with non-frontal lesions, and healthy control subjects. Individuals with psychopathy and those with OFC lesions were impaired on the “affective ToM” conditions, but not in “cognitive ToM” conditions, compared with the control groups. This study suggests that mentalizing impairments in psychopathy resembles remarkably that seen in participants with lesions of the frontal lobe, particularly with OFC damage, and provides support for the notion of amygdala-OFC dysfunction in psychopathy (Blair, 2007; Finger, Marsh, Mitchell, Reid, Sims, Budhani, et al., 2008). However, it should be noted that regardless of the similarities between acquired frontal lesions and developmental psychopathy, comparison between these groups should be treated with caution as essential differences exist between these individuals. Importantly, while both groups may demonstrate reactive aggression, instrumental aggression is typically reported in developmental psychopathy, but rarely reported after OFC damage (Mitchell, Avny & Blair, 2006).

Frontotemporal dementia

Frontotemporal dementia (FTD) is a progressive neurodegenerative syndrome with diverse clinical presentations. Among the most prominent features are progressive aphasia and bizarre affect with a “personality change.” These may include disinhibition and impulsivity, distractibility and impersistence, and perseverative behavior, in addition to social deficits such as lack of empathy, emotional unconcern, apathy, and irritability (for a review, see Grossman, 2002). Shany-Ur and colleagues (Shany-Ur, Poorzand, Grossman, Growdon, Jang, Ketelle, et al., 2012) investigated whether face-to-face testing of comprehending insincere communication would effectively discriminate among different neurodegenerative disease patients. The authors examined the ability to comprehend lies and sarcasm from a third-person perspective, using contextual cues, in 102 patients with either FTD, Alzheimer’s disease, progressive supranuclear palsy (PSP) or vascular cognitive impairment, and 77 healthy older adults. Participants answered questions about videos depicting social interactions involving deceptive, sarcastic, or sincere speech using the Awareness of Social Inference Test, which assesses poor understanding of emotional expressions and difficulty integrating contextual information that is part of normal social encounters (McDonald, Flanagan, Rollins, 2002). All subjects equally understood sincere remarks, but FTD patients displayed impaired comprehension of lies and sarcasm compared with normal controls. In other groups, impairment was not disease-specific, but was proportionate to general cognitive impairment. Analysis of the task components revealed that only FTD patients were impaired on perspective taking and emotion reading elements and that both FTD patients and PSP patients had impaired ability to represent others opinions and intentions (i.e. ToM). Moreover, test performance correlated with informants’ ratings of subjects’ empathy, perspective taking and neuropsychiatric symptoms in everyday life. All patient groups exhibited some deficiencies in these complex social communication tasks, which require multiple cognitive and emotional processes. However, FTD patients showed uniquely focal and severe impairments at every level of ToM and emotion reading skills, showing an inability to identify even obvious examples of deception and sarcasm. These results suggest that FTD may target a specific neural network necessary for perceiving social salience and social outcomes (Shany-Ur et al., 2012).

Interactions between the two empathy systems

The differentiation between these two systems enabling empathy suggests that under normal circumstances every interaction with another may trigger independently both an emotional response

(emotional empathy) as well as cognitive evaluation of his state of mind or perspective (cognitive empathy). The other’s emotions are “shared” through brain areas involved in resonance or simulation, such as the human MNS. In addition, the ability to accurately infer the other’s perspective and imagine their state of mind is activated, requiring self-other decoding and ToM abilities. Both functional neuroimaging and lesion studies in humans indicate that the vmPFC plays a crucial role in the network performing cognitive empathic function. This system is phylogenetically younger and is unique to primates and human adults.

Although both emotional and cognitive components of empathy may operate partly autonomously, it is likely that every empathic response will evoke both components to some extent, depending on the social context. Zaki and colleagues showed that empathically accurate, as compared with inaccurate, judgments depended both on the activation of structures within the human MNS, thought to be involved in emotional empathy, and on the activation of regions implicated in mental state attribution, or cognitive empathy, such as the medial prefrontal cortex. These data demonstrate that activity in these two sets of brain regions tracks with the accuracy of attributions made about another’s internal emotional state (Zaki, Bolger, & Ochsner, 2009). Moreover, in a second study, the authors manipulated the degree to which one could use non-verbal and contextual cues in order to infer about the target’s emotions. They found that conditions in which biasing of neural activity was seen toward the MNS, tracked with perceivers’ behavioral reliance on non-verbal cues; In contrast, conditions in which biasing was seen toward what they called a “mental state attribution system”, including the TPJ and mPFC, was tracked with perceivers’ reliance on contextual cues when drawing inferences about targets’ emotions. Future studies may further this focus on the interactions between the two systems and the different conditions that may affect the activation of each. Different variables such as the level of emotions involved, the past experiences of the empathizer, gender, relationship with the protagonist and the perceived similarity between

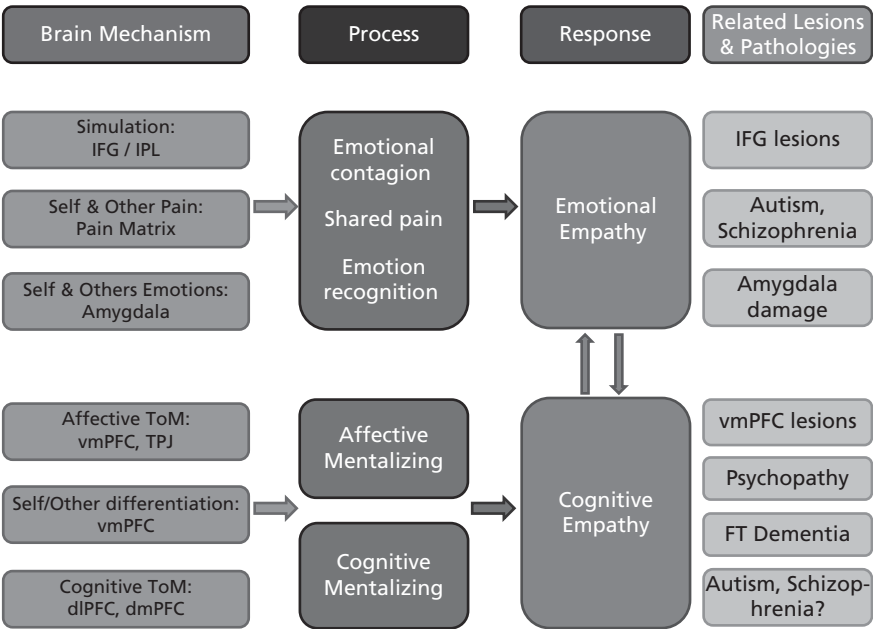


Figure 11.3 Summary and proposed model.

the individual and the protagonist may differentially activate the emotional and the cognitive systems. Research of complex conditions, such as autism and schizophrenia, which are characterized by (among others) their deficits in social cognition and empathic abilities, may highly benefit from the progress in this field. Exploring these questions using a combination of several research tools such as lesion studies, neuroimaging, electrophysiology, genetics and animal research may prove to be essential in characterizing the relationship between these two systems and the conditions in which each system is activated. See Figure 11.3 for a proposed model.

References

- Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences* 1191: 42–61.
- Adolphs, R., Spezio, M. L., Parlier, M., & Piven, J. (2008). Distinct face processing strategies in parents of autistic children. *Current Biology* 18: 1090–3.
- Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* 372: 669–72.
- Adolphs, R., Tranel, D., Hamann, S., Young, A. W., Calder, A. J., Phelps, E. A., Anderson, A., Lee, G. P., & Damasio, A. R. (1999). Recognition of facial emotion in nine subjects with bilateral amygdala damage. *Neuropsychologia* 37: 1111–17.
- Arbib, M. A., & Mundhenk T. N. (2005) Schizophrenia and the mirror-neuron system: an essay. *Neuropsychologia* 43: 268–80.
- Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. In: S. Baron-Cohen, H. Tager-Flusberg & D. J. Cohen (Eds), *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience* (pp. 1–20). Oxford: Oxford University Press.
- Blair R. J. (2007). Dysfunctions of medial and lateral orbitofrontal cortex in psychopathy. *Annals of the New York Academy of Sciences* 1121: 461–79.
- Blair R. J., & Cipolotti L. (2000). Impaired social response reversal. A case of “acquired sociopathy”. *Brain* 123:1122–41.
- Brothers, L., & Ring, B. (1992). A neuroethological framework for the representation of minds. *Journal of Cognitive Neuroscience* 4: 107–18.
- Buccino, G., & Amore M. (2008). Mirror neurons and the understanding of behavioural symptoms in psychiatric disorders. *Current Opinion in Psychiatry* 21: 281–5.
- Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., et al. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience* 13(2): 400–4.
- Burns, J. (2006). The social brain hypothesis of schizophrenia. *World Psychiatry* 5: 77–81.
- Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Science* 12: 187–92.
- Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs in Social Research in Child Development* 63: 176.
- Carrington, S. J., Bailey, A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping* 30: 2313–35.
- Carruthers, P., & Smith, P. K. (Eds). (1996). *Theories of Theories of Mind*: Cambridge University Press.
- Chakrabarti, B., Bullmore, E. T., & Baron-Cohen, S. (2006). Empathizing with basic emotions: Common and discrete neural substrates. *Social Neuroscience* 1(3–4): 364–84.
- Cheng, Y., Lee, P. L., Yang, C. Y., Lin, C. P., Hung, D., & Decety, J. (2008a). Gender differences in the mu rhythm of the human mirror-neuron system. *PLoS ONE* 3(5): e2113.
- Cheng, Y., Yang, C-Y., Lin, C-P., Lee, P-L., & Decety, J. (2008b). The perception of pain in others suppresses somatosensory oscillations: A magnetoencephalography study. *NeuroImage* 40(4): 1833–40.

- Cochin, S., Barthelemy, C., Roux, S., & Martineau, J. (1999). Observation and execution of movement: similarities demonstrated by quantified electroencephalography. *European Journal of Neuroscience* 11(5): 1839–42.
- Cochin, S., Barthelemy, C., Lejeune, B., Roux, S., & Martineau, J. (1998). Perception of motion and qEEG activity in human adults. *Electroencephalography and Clinical Neurophysiology* 107(4): 287–95.
- Churchland, P. N. (1988). *Matter and Consciousness*. Cambridge: MIT Press.
- Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology* 10, 85.
- Dalton, K. M., Nacewicz, B. M., Alexander, A. L., & Davidson, R. J. (2007). Gaze-fixation, brain activation, and amygdala volume in unaffected siblings of individuals with autism. *Biological Psychiatry* 61: 512–20.
- Dapretto, M., Davies M. S., Pfeifer J. H., Scott A. A., Sigman M., Bookheimer S. Y., et al. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience*, 9, 28–30.
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology* 44(1): 113–26.
- Davies, M., & Stone, T. (Eds). (1995). *Mental Simulation*. Oxford: Basil Blackwell.
- Das, P., Kemp, A. H., Flynn, G., Harris, A. W., Liddell, B. J., Whitford, T. J., et al. (2007). Functional disconnections in the direct and indirect amygdala pathways for fear processing in schizophrenia. *Schizophrenia Research* 90: 284–94.
- de Waal, F. B. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual Review of Psychology* 59: 279–300.
- Decety, J., Chaminade, T., Grezes, J., & Meltzoff, A. N. (2002). A PET exploration of the neural mechanisms involved in reciprocal imitation. *NeuroImage* 15(1): 265–72.
- Decety J., Yang C. Y., & Cheng Y. (2010). Physicians down-regulate their pain empathy response: An event-related brain potential study. *NeuroImage* 50: 1676–82.
- di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research* 91(1): 176–80.
- Enticott P. G., Hoy K. E., Herring S. E., Johnston P. J., Daskalakis Z. J., & Fitzgerald P. B. (2008). Reduced motor facilitation during action observation in schizophrenia: A mirror neuron deficit? *Schizophrenia Research* 102:116–21.
- Eslinger, P. J. (1998) Neurological and neuropsychological bases of empathy. *European Neurology* 39: 193–9.
- Eslinger, P. J., Flaherty-Craig, C. V., & Benton, A. L. (2004). Developmental outcomes after early prefrontal cortex damage. *Brain and Cognition* 55(1): 84–103.
- Finger E. C., Marsh A. A., Mitchell D. G. V., Reid M. E., Sims C., Budhani S., et al. (2008). Abnormal ventromedial prefrontal cortex function in children with psychopathic traits during reversal learning. *Archives of General Psychiatry* 65, 586–94.
- Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1-Level 2 distinction. *Developmental Psychology* 17, 99–103.
- Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance—reality distinction. *Cognitive Psychology* 15(1): 95–120.
- Fodor, J. A. (1987). *Psychosemantics*. Cambridge: MIT Press.
- Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308(5722): 662–7.
- Frith, C. D., & Frith, U. (1999). Interacting minds—a biological basis. *Science* 286(5445): 1692–5.
- Frith C. D., & Singer T. 2008. The role of social cognition in decision making. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences* 363: 3875–86.
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119(Pt 2): 593–609.

- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* 2(12): 493–501.
- Gastaut, H. (1952). Etude électrocorticographique de la réactivité des rythmes rolandiques. *Revue Neurologique* 87(2): 176–82.
- Gastaut, H. J., & Bert, J. (1954). EEG changes during cinematographic presentation. *Electroencephalography and Clinical Neurophysiology* 6: 433–44.
- Goldman, A. (1989). Interpretation psychologized *Mind Language* 4, 161–85.
- Gopnik, A. (1993). How we know our minds: the illusion of first-person knowledge of intentionality. *Behaviour & Brain Sciences* 16: 1–14.
- Gordon, R. (1986). Folk psychology as simulation. *Mind Language* 1, 158–71.
- Grafton, S. T., Arbib, M. A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations in humans by positron emission tomography. 2. Observation compared with imagination. *Experimental Brain Research* 112(1): 103–11.
- Grezes, J., Armony, J. L., Rowe, J., & Passingham, R. E. (2003). Activations related to “mirror” and “canonical” neurones in the human brain: An fMRI study. *NeuroImage* 18(4): 928–37.
- Grezes, J., Costes, N., & Decety, J. (1998). Top down effect of strategy on the perception of human biological motion: a PET investigation. *Cognitive Neuropsychology*, 15: 553–82.
- Grossman, M. (2002). Frontotemporal dementia: a review. *Journal of the International Neuropsychology Society* 8, 566–83.
- Gur, R. E., Calkins, M. E., Gur, R. C., Horan, W. P., Nuechterlein, K. H., Seidman, L. J., Stone, W. S. (2007). The Consortium on the Genetics of Schizophrenia: neurocognitive endophenotypes. *Schizophrenia Bulletin* 33(1), 49–68.
- Gur, R. E., Loughhead, J., Kohler, C. G., Elliott, M., Lesko, K., Ruparel, K., Wolf, D. H., Bilker W. B., & Gur, R. C. (2007). Limbic activation associated with misidentification of fearful faces and flat affect in schizophrenia. *Archives of General Psychiatry* 64, 1356–66.
- Harlow, J. M. (1868). Recovery from the passage of an iron bar through the head. *Publications of the Massachusetts Medical Society* 2: 327–47.
- Heal, J. (Ed.). (1986). *Replication and functionalism*. Cambridge: Cambridge University Press.
- Hooker, H. I., Bruce, L., Lincoln, S. H., Fisher, M., & Vinogradov, S. (2011) Theory of mind skills are related to gray matter volume in the ventromedial prefrontal cortex in schizophrenia. *Biological Psychiatry* 70(12): 1169–78.
- Hynes, C. A., Baird, A. A., & Grafton, S. T. (2006). Differential role of the orbital frontal lobe in emotional vs. cognitive perspective-taking. *Neuropsychologia* 44, 374–83.
- Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotto, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science* 286(5449): 2526–8.
- Jabbi, M., Swart, M., Keysers, C. (2007). Empathy for positive and negative emotions in the gustatory cortex. *NeuroImage* 34: 1744–53.
- Johnson, S. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358(1431): 549.
- Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., Brand, M., Shamay-Tsoory, S., Onur, O. A., & Kessler, J. (2010). Dissociating cognitive from affective theory of mind: a TMS study. *Cortex* 46, 769–80.
- Keysers, C., & Gazzola, V. (2007). Integrating simulation and theory of mind: from self to social cognition. *Trends in Cognitive Sciences* 11(5): 194–6.
- Keysers, C., Kaas, J. H., & Gazzola, V. (2010). Somatosensation in social perception. *National Review of Neuroscience* 11: 417–28.
- Kosslyn, S. (1978). Measuring the visual angle of the mind’s eye. *Cognitive Psychology* 10, 356–89.
- Lamm, C., Meltzoff, A. N., & Decety, J. (2010). How do we empathize with someone who is not like us? A functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience* 22, 362–76.

- Langford D. J., Crager S. E., Shehzad Z., Smith S. B., Sotocinal S. G., Levenstadt J. S. Chanda M. L., Levitin D. J., & Mogil J. S. (2006). Social modulation of pain as evidence for empathy in mice. *Science* 312, 1967–70.
- Larson, C. L., Schaefer, H. S., Siegle, G. J., Jackson, C. A., Anderle, M. J., & Davidson, R. J. (2006). Fear is fast in phobic individuals: amygdala activation in response to fear-relevant stimuli. *Biological Psychiatry* 60, 410–17.
- Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* 50, 249–58.
- Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Sadek, S. A., Pasco, G., Wheelwright, S. J., Suckling, J., & Baron-Cohen, S. (2010). Atypical neural self-representation in autism. *Brain* 133, 611–24.
- Martineau, J., Schmitz, C., Assaiante, C., Blanc, R., & Barthelemy, C. (2004). Impairment of a cortical event-related desynchronization during a bimanual load-lifting task in children with autistic disorder. *Neuroscience Letters* 367(3): 298–303.
- McDonald, S., Flanagan, S., & Rollins, J. (2002). *The Awareness of Social Inference*. TestSuffolk: Thames Valley Test Co., Ltd.
- Mitchell, J. P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society London, B, Biological Sciences* 364, 1309–16.
- Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006) Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron* 50(4): 655–63.
- Mitchell, D., Avny, S., & Blair, R. J. (2006). Divergent patterns of aggressive and neurocognitive characteristics in acquired versus developmental psychopathy. *Neurocase* 12(3): 164–78
- Muthukumaraswamy, S. D., Johnson, B. W., & McNair, N. A. (2004). Mu rhythm modulation during observation of an object-directed grasp. *Brain Research in Cognitive Brain Research* 19(2): 195–201.
- Niedenthal, P. (2007). Embodying emotion. *Science* 316(5827): 1002.
- Nummenmaa, L., Hirvonen, J., Parkkola, R., & Hietanen, J. K. (2008). Is emotional contagion special? An fMRI study on neural systems for affective and cognitive empathy. *NeuroImage* 43, 571–80.
- Oberman, L. M., Hubbard, E. M., McCleery, J. P., Altschuler, E. L., Ramachandran, V. S., & Pineda, J. A. (2005). EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Brain Research: Cognitive Brain Research* 24(2): 190–8.
- Oberman, L. M., Pineda, J. A., & Ramachandran, V. S. (2007). The human mirror neuron system: A link between action observation and social skills. *Social Cognitive Affective Neuroscience* 2(1): 62–6.
- Oberman, L. M., Ramachandran, V. S., & Pineda, J. A. (2008). Modulation of mu suppression in children with autism spectrum disorders in response to familiar or unfamiliar stimuli: The mirror neuron hypothesis. *Neuropsychologia* 46(5): 1558–65.
- Park, I. H., Park, H. J., Chun, J. W., Kim, E. Y., & Kim, J. J. (2008). Dysfunctional modulation of emotional interference in the medial prefrontal cortex in patients with schizophrenia. *Neuroscience Letters* 440, 119–24.
- Perry, A., Bentin, S., Ben-Ami Bartal, I., Lamm, C., & Decety, J. (2010). “Feeling” the pain of those who are different from us—modulation of EEG in the mu/alpha range. *Cognitive, Affective and Behavioral Neuroscience* 10, 493–504.
- Perry, A., Troje, N. F., & Bentin, S. (2010) Exploring motor system contributions to the perception of social information: Evidence from EEG activity in the mu/alpha frequency range. *Social Neuroscience* 5(3): 272–84.
- Pezawas, L., Meyer-Lindenberg, A., Drabant, E. M., Verchinski, B. A., Munoz, K. E., Kolachana, B. S., Egan, M. F., Mattay, V. S., Hariri, A. R., & Weinberger, D. R. (2005). 5-HTTLPR polymorphism impacts human cingulate-amygdala interactions: a genetic susceptibility mechanism for depression. *National Neuroscience* 8: 828–34.
- Pineda, J. A. (2005). The functional significance of mu rhythms: translating “seeing” and “hearing” into “doing”. *Brain Research and Brain Research Reviews* 50(1): 57–68.

- Premack, D., & Woodruff, G. (1978). Chimpanzee problem-solving: a test for comprehension. *Science* 202(4367): 532–5.
- Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences* 25(1): 1–20.
- Raymaekers, R., Wiersema, J., & Roeyers, H. (2009). EEG study of the mirror neuron system in children with high functioning autism. *Brain Research* 1304: 113–21.
- Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research and Cognitive Brain Research* 3(2): 131–41.
- Rozzi, S., Ferrari, P. F., Bonini, L., Rizzolatti, G., & Fogassi, L. (2008). Functional organization of inferior parietal lobule convexity in the macaque monkey: electrophysiological characterization of motor, sensory and mirror responses and their correlation with cytoarchitectonic areas. *European Journal of Neuroscience* 28(8): 1569–88.
- Saarela, M. V., Hlushchuk, Y., Williams, A. C., Schurmann, M., Kalso, E., & Hari, R. (2007). The compassionate brain: humans detect intensity of pain from another's face. *Cerebral Cortex* 17: 230–7.
- Schulte-Ruther, M., Markowitsch, H. J., Fink, G. R., & Piefke, M. (2007). Mirror neuron and theory of mind mechanisms involved in face-to-face interactions: a functional magnetic resonance imaging approach to empathy. *Journal of Cognitive Neuroscience* 19, 1354–72.
- Schulte-Ruther, M., Greimel, E., Markowitsch, H. J., Kamp-Becker, I., Remschmidt, H., Fink, G. R., & Piefke, M. (2010). Dysfunctions in brain networks supporting empathy: An fMRI study in adults with autism spectrum disorders. *Social Neuroscience* 6: 1–21.
- Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. *Neuropsychologia* 45: 3054–67.
- Shamay-Tsoory, S. G., Aharon-Peretz, J., & Levkovitz, Y. (2007) The neuroanatomical basis of affective mentalizing in schizophrenia: comparison of patients with schizophrenia and patients with localized prefrontal lesions. *Schizophrenia Research* 90, 274–83.
- Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus vs. ventromedial prefrontal lesions. *Brain* 132, 617–27.
- Shamay-Tsoory, S. G., Harari, H., Aharon-Peretz, J., & Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex* 46(5): 668–77.
- Shamay-Tsoory, S. G., Shur, S., Barcai-Goodman, L., Medlovich, S., Harari, H., & Levkovitz, Y. (2007). Dissociation of cognitive from affective components of theory of mind in schizophrenia. *Psychiatry Research* 149(1–3): 11–23.
- Shamay-Tsoory, S. G., Tibi-Elhanani, Y. & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective, but not cognitive theory of mind stories. *Social Neuroscience* 1(3–4): 149–66.
- Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., Goldsher, D., & Aharon-Peretz, J. (2005). Impaired “affective theory of mind” is associated with right ventromedial prefrontal damage. *Cognitive and behavioral neurology* 18(1): 55–67.
- Shamay-Tsoory, S. G., Tomer, R., Goldsher, D., Berger, B. D., & Aharon-Peretz, J. (2004). Impairment in cognitive and affective empathy in patients with brain lesions: anatomical and cognitive correlates. *Journal of Clinical Experimental Neuropsychology* 8, 1113–27.
- Shany-Ur, T., Poorzand, P., Grossman, S., Growdon, M. E., Jang, J. Y., Ketelle, R. S., Miller, B. L., & Rankin, K. P. (2012). Comprehension of insincere communication in neurodegenerative disease: Lies, sarcasm, and theory of mind. *Cortex* 48(10): 1329–41.
- Simner, M. L. (1971) Newborn's response to the cry of another infant. *Developmental Psychology* 5: 136–50.
- Singh, F., Pineda, J., & Cadenhead, K. S. (2011). Association of impaired EEG mu wave suppression, negative symptoms and social functioning in biological motion processing in first episode of psychosis. *Schizophrenia Research* 130, 182–6.

- Spezio, M. L., Adolphs, R., Hurley, R. S., & Piven, J. (2007). Abnormal use of facial information in high-functioning autism. *Journal of Autism and Developmental Disorders* 37: 929–39.
- Stich, S. A. N., & Nichols, S. (1992). Folk psychology: simulation vs. tacit theory. *Mind and Language* 7: 29–65.
- Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience* 10(5): 640–56.
- Tranel, D., Bechara, A., & Denburg, N. L. (2002). Asymmetric functional roles of right and left ventromedial prefrontal cortices in social conduct, decision-making, and emotional processing. *Cortex* 38: 589–612.
- Uddin, L. Q., Davies, M. S., Scott, A. A., Zaidel, E., Bookheimer, S. Y., Iacoboni, M., et al. (2008). Neural basis of self and other representation in autism: An FMRI study of self-face recognition. *PLoS ONE* 3: 3526.
- Van Overwalle F., & Baetens K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *NeuroImage* 48(3): 564–84.
- Völlm, B. A., Taylor, A. N. W., Richardson, P., Corcoran, R., Stirling, J., McKie, S., et al. (2006). Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a non-verbal task. *NeuroImage* 29: 90–8.
- Wellman, H. (1990). *The Child's Theory of Mind*. Cambridge: MIT Press.
- Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron* 40(3): 655–64.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13(1): 103–28.
- Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion* 9: 478–87.

Chapter 12

Empathy and the brain

Cade McCall and Tania Singer

Feeling what others feel is basic to human social life. We wince when we see someone's finger sliced by a razor, when we see that person's face twist in pain, or even when we read about the event. Both physical cues and our imaginations are enough for us to infer and experience the affective states of others. These abilities have clear functional benefits, allowing us to learn from others' pain and to offer help and support when they're needed. While empathy is closely related to mentalizing about others' thoughts and intentions, sharing feelings is distinct from reading minds. In recent years, social neuroscience has made major strides in understanding empathy. Research on its neural representations and modulation has produced a complex picture. There is no single brain region underlying empathy, but a variety of networks that work together to produce (and prevent) vicarious feeling. Significant questions remain, particularly regarding the different domains of empathic experience, its developmental trajectories, and the translation of shared feelings into behavior. This chapter provides an overview of this work and highlights possible new directions for research.

Defining empathy

Definitions of empathy vary widely in their focus and breadth. Based heavily on groundbreaking work in psychology (Batson, 2009b; Eisenberg & Fabes, 1990; Wispe, 1986), social neuroscientists have honed in on a relatively specific construct for the purposes of research. One definition of empathy recently proposed by neuroscientists, for example, has four key components (de Vignemont & Singer, 2006; Decety & Jackson, 2004; Singer & Lamm, 2009). First, empathy refers to an affective state. Secondly, that state is elicited by the inference or imagination of another person's state. Thirdly, that state is isomorphic with the other person's state. Fourthly, the empathizer knows that the other person is the source of the state. In other words, empathy is the experience of vicariously feeling what another person is feeling without confounding the feeling with one's own direct experience (see Figure 12.1).

This definition distinguishes empathy from related phenomena. While mentalizing or cognitive perspective-taking may help us infer another person's affective state, it does not necessarily produce an affective state in ourselves. For example, mentalizing might produce the inference, "I see him smiling so he must be happy," while empathizing would produce the experience, "I am happy because he's happy." In gross terms, mentalizing represents more "cold" cognitive analysis of the scene and empathy the "warm" experiential response. Nevertheless, while these two constructs may be distinct on paper, mentalizing and empathy are closely related in mental life (Jackson, Brunet, Meltzoff, & Decety, 2006). For example, mentalizing plays a key part in providing the cues necessary to trigger empathic reactions. Conversely, empathic experience likely contributes to our mentalizing abilities by teaching us the meanings of specific affective cues.

Emotional contagion is another closely related phenomenon. In emotional contagion one "catches" the affective state of another person, but without awareness of the state's source

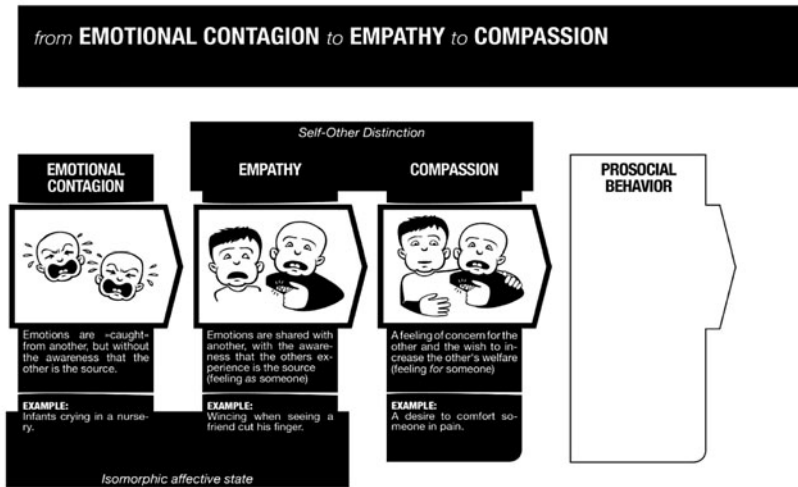


Figure 12.1 The conceptual relationship between emotional contagion, empathy, and compassion.

(de Vignemont & Singer, 2006; Hatfield, Cacioppo, & Rapson, 1994). The automatic spread of panic through a crowd or the collective crying among babies in a nursery are examples. The critical difference between empathy and emotional contagion is that empathy maintains a self-other distinction. In other words, it is clear to the empathizer that the target is the source of the affective state. Nevertheless, emotional contagion is likely a developmental precursor for empathy (Hatfield, Rapson, & Le, 2009; Singer & Lamm, 2009). Moreover, the mechanisms responsible for emotional contagion may also function in full-blown empathy, but with the fine-tuning of the self-other distinction (Decety & Jackson, 2004).

Sympathy and pity are also affective responses to another person's state, but without the isomorphic quality of empathy (Batson, 2009b). For example, "The fact that he's angry makes me sad," or "I'm happy that he's comfortable." Note that neither sympathy nor pity involves a clear element of emotional contagion. They are, however, likely to involve some degree of mentalizing as the empathizer uses various cues and beliefs about the other person's goals and experiences to infer an affective state. In terms of affect, sympathy is less direct than empathy (de Vignemont & Singer, 2006; Singer & Lamm, 2009); it involves feeling **for** someone, not feeling **as** someone (Batson, 2009b).

The term "compassion" is often used interchangeably with empathy. There is, however, an important distinction based on motivation and behavior. Compassion is characterized by the motivation to alleviate the distress of another (Baumeister & Vohs, 2007). Although empathy may allow a compassionate individual to know when and how to act (Batson, 2009a; Eisenberg, 2000), empathy does not always result in compassion. In fact, empathic distress may lead the empathizer to avoid the target individual (Batson, Fultz, & Schoenrade, 1987; Eisenberg & Fabes, 1990). Conversely, compassion is an approach-oriented response to the affective state of the other. It represents the prosocial consequences of empathic experience. One can also imagine antisocial responses that rely upon empathy (Singer & Lamm, 2009). Take, for example, a torturer who is uniquely skilled at knowing what will cause pain to his victims. The ability to experience vicarious pain would help him know how to hurt others.

The four parts of this definition of empathy provide a relatively well circumscribed territory for the study of this phenomenon. When it comes to neural underpinnings, we expect empathic

neural responses to represent affect (i) and to do so in a way that reflects the specific affective state (iii) of the empathized other (ii). At the same time, we expect distinctions between self and other representations indicative of the fact that the empathizer's feeling is vicarious and not direct (iv).

Neural representations of empathic states

In recent years, researchers have created a novel set of experimental paradigms to study the neuroscience of empathy. Initially this work was built on the premise that if empathy represents shared affect, then the neural representations of those vicarious states should show at least some overlap with self-generated representations of that same affective state (Avenanti, Buetti, Galati, & Aglioti, 2005; Botvinick, Jha, Bylsma, Fabian, Solomon, & Prkachin, 2005; Jackson, Meltzoff, & Decety, 2005; Keysers, Wicker, Gazzola, Anton, Fogassi, & Gallese, 2004; Morrison, Lloyd, Di Pellegrino, & Roberts, 2004; Singer, Seymour, O'Doherty, Kaube, Dolan, & Frith, 2004; Wicker, Keysers, Plailly, Royet, Gallese, & Rizzolatti, 2003). This "shared network hypothesis" emerged, in part, out of evidence for shared cognitive and neural representations of action and perception (Preston & de Waal, 2002; Gallese & Goldman, 1998; Prinz, 1997; Prinz, 2005). Research on the cognition mechanisms underlying action, for example, has consistently demonstrated that watching another person executing an action can interfere with the planning and execution of an incongruent action, and can facilitate the planning and execution of a congruent action. These types of findings suggest that a common coding exists for one's own actions and the perceived actions of others (Prinz, 2005; Prinz, 1997). This evidence was further bolstered by the discovery of mirror neurons in Macaque monkeys, neurons that respond to both action and the perception of action (di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti & Craighero, 2004). Cognitive neuroscience, in turn, has demonstrated overlaps between regions representing one's own and others actions (Buccino, Binkofski, Fink, Fadiga, Fogassi, Gallese, et al., 2001; Jeannerod, 2001). Together these data led researchers to suggest that social cognition is built on the automatic simulation of others' behaviors (Gallese & Goldman, 1998; Keysers & Gazzola, 2007; Rizzolatti, Fogasi, & Gallese, 2001). The brain perceives others' actions and through their simulation infers the meanings of those actions.

Following this line of reasoning, the shared network hypothesis of empathy suggests that we understand others' affective states by recruiting the same networks that represent our own affective states. Direct and vicarious feeling rely on similar mechanisms.

Empathy for pain

While the shared network hypothesis has been tested in several affective domains, empathy for pain has been a particularly fruitful target. Pain lends itself to this line of research because it is easily manipulated and depicted within the laboratory. Both pain and empathizing for another person's pain are common and salient experiences. Perhaps most importantly, the "pain matrix", or networks responsible for representing pain (Apkarian, Bushnell, Treede, & Zubieta, 2005; Derbyshire, 2000; Peyron, Laurent, & Garcia-Larrea, 2000), is relatively well understood. As a consequence, researchers can make clear predictions about locations of overlapping representation and can theorize about the specific features of pain that are vicariously represented.

Social neuroscientists have used two distinct methods to manipulate and measure empathy for pain in the laboratory: picture-based and cue-based paradigms. In picture-based paradigms (e.g. Jackson et al., 2005), participants view pictures or videos depicting painful situations. For example, they might see a q-tip stroking a hand during a non-painful trial or a needle puncturing a hand during a painful trial (Lamm, Meltzoff, & Decety, 2010). Alternatively, the images can depict the

face of an individual as he or she experiences pain (Saarela, Hlushchuk, Williams, Schurmann, Kalso, & Hari, 2007). These studies allow researchers to measure neural responses while manipulating the nature of the vicarious stimulus, the location of that stimulus on the target's body, and the affective response of the pained individual.

Cue-based paradigms, on the other hand, use actual people instead of images as stimuli (Singer et al., 2004). During these experiments, multiple participants both receive and witness the delivery of painful stimuli (i.e. electric shocks to the hand). On each trial in such a study, a cue indicates (a) whether or not the stimulus for that trial will be painful, and (b) the recipient of that stimulus (i.e. self or other). Because this paradigm uses arbitrarily assigned cues to indicate trial type, any responses that emerge during other-recipient trials are entirely cue-triggered and cannot be caused by emotional contagion. They cannot be driven by simply seeing the recipient's body or by expressions of affect. In other words, empathic responses in these studies are the consequence of knowing the other person is in pain and imagining that state, not in perceiving the other person's actual response to that pain. The other important feature of this paradigm is that it includes both direct and vicarious pain trials; participants both experience and witness experience pain. As a consequence, researchers can perform a direct, within-subject comparison between a participant's own pain and his or her reaction to another's pain (see also Corradi-Dell'Acqua, Hofstetter, & Vuilleumier, 2011).

According to shared network hypotheses, empathy for another person's pain should activate components of the pain matrix. This activation should emerge when contrasting neural activity during trials depicting painful vs. non-painful trials. In experiments that include direct pain trials, one should also find overlap between self and other pain representations. Although cue-based and picture-based paradigms furnish distinct patterns of data, recent meta-analyses provide strong evidence for a core network for empathy for pain. One image-based meta-analysis representing 9 separate studies (Lamm, Decety, & Singer, 2011) and two coordinate-based meta-analyses representing 32 (Lamm et al., 2011) and 40 studies (Fan, Duncan, Greck, & Northoff, 2011) found significant bilateral anterior insula (AI), dorsal anterior cingulate cortex (ACC) and anterior midcingulate cortex (aMCC) activity during empathy for pain across a variety of experiments conducted by different research groups (Figure 12.2, Panel 1). Critically, these areas overlap with areas that emerged in a meta-analysis of activity during the direct experience of pain (Figure 12.2, Panel 4; Lamm et al., 2011).

Participant self-reports of empathic states and traits corroborate the role of these areas during the representation of vicarious pain. Activity in the ACC and left AI during other pain trials correlate (Singer et al., 2004; Singer, Seymour, O'Doherty, Stephan, Dolan, & Frith, 2006; Jackson et al., 2005; Lamm, Batson, & Decety, 2007a) with dispositional measures of empathy such as the Balanced Emotional Empathy Scale (Mehrabian & Epstein, 1972) and the Empathic Concern subscale of the Interpersonal Reactivity Index (Davis, 1983). Similar findings have been reported with the Empathy Quotient (Baron-Cohen & Wheelwright, 2004) and measures of emotional contagion (Lamm et al., 2007a; Doherty, 1997); see also (Jabbi, Swart, & Keysers, 2007). Reports of perceived target pain intensity or unpleasantness on a trial-by-trial basis also correlate with ACC and AI activity during those trials (Jackson et al., 2005; Saarela et al., 2007; Singer, Snozzi, Bird, Petrovic, Silani, Heinrichs, et al., 2008; Lamm, Nusbaum, Meltzoff, & Decety, 2007b; Cheng, Lin, Liu, Hsu, Lim, Hung, et al., 2007).

The core regions found across studies on empathy for pain map onto some, but not all, of the pain matrix. Here, qualitative distinctions between features of painful experience are critical. Specifically, the pain matrix can be divided into regions that represent sensory discriminative (the location of the pain, the quality of the nociceptive input, etc.) vs. affective and motivational components of pain (the experience of unpleasantness, avoidance motivation, etc.; Apkarian

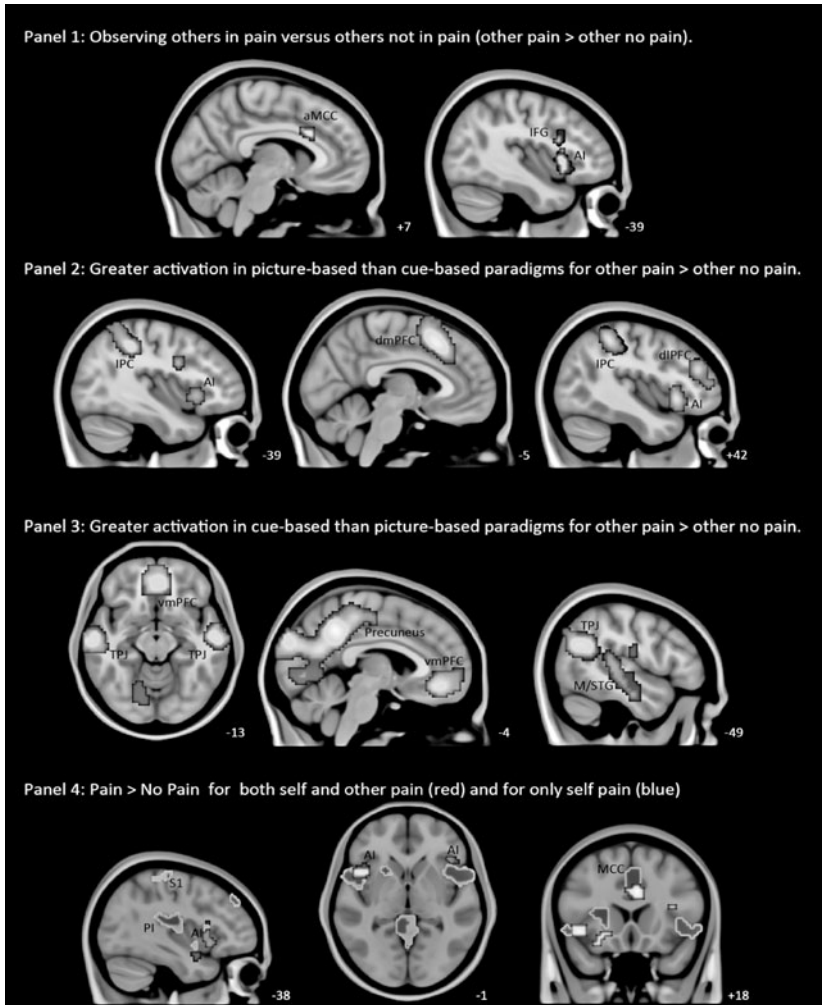


Figure 12.2 Results of a meta-analysis of empathy for pain studies (Lamm et al., 2011). The areas highlighted in Panel 1 showed significantly more activity when participants observed others in pain (as compared with trials in which they observed others not in pain). Panel 1 includes data from both cue-based and picture-based paradigms. Panel 2 depicts regions that showed higher activations for this contrast in picture-based paradigms. Conversely, Panel 3 depicts regions that showed higher activation for this contrast in cue-based paradigms. Panel 4 depicts areas that were common to both the experience of pain and the observation of others in pain (the bright spectrum) as well as areas that were unique to the direct experience of pain (the dark spectrum). See also Plate 5.

Analyses from Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage* 54(3): 2492–502. © 2011, Elsevier.

et al., 2005; Peyron et al., 2000). Somatosensory cortices (S1 and S2) and the posterior insula are implicated in the sensory discriminative components of pain (Apkarian et al., 2005; Maihöfner, Herzner, & Handwerker, 2006). In line with their role in pain localization, both the posterior insula and somatosensory cortices show activity contralateral to the location of the painful stimulus on

the body (Bingel, Quante, Knab, Bromm, Weiller, & Buchel, 2003; Brooks, Nurmikko, Bimson, Singh, & Roberts, 2002). Moreover, damage to the SI and SII (but not the AI or ACC) selectively impairs the ability to discern the quality and localization of pain without impairing the ability to experience an unpleasant feeling that varies with stimulus intensity (Ploner, Freund, & Schnitzler, 1999). Activation in S2, furthermore, varies with self reports of the sensory-discrimination, but not the other components of the pain (Maihöfner et al., 2006; Melzack, 1975).

On the other hand, the areas active across empathy for pain paradigms (the AI, dACC, and aMCC) have been implicated in pain's abstract and affective features (Price, 2000). AI activity, for example, varies not only with the level of noxious input, but additionally as a function of self-reported intensity (Kong, Gollub, Polich, Kirsch, Laviolette, Vangel, et al., 2008). Similarly, activity in the ACC correlates with self-reported unpleasantness of pain (Rainville, Duncan, Price, Carrier, & Bushnell, 1997). Not surprisingly then, vicarious pain involves feelings of discomfort and aversion without necessarily involving the more sensory-specific qualities of the stimulus. This is not to say, however, that areas associated with the sensory components of pain are never involved in empathic experience. Instead, it appears that different networks lead to the elicitation of the empathy and that the empathic experience of pain converges on this core network (Lamm et al., 2011).

Different routes to empathic experience

Comparisons between the results of cue-based and picture-based empathy for pain paradigms reveal important differences in patterns of activation. These differences suggest that the brain elicits empathic states through different computational routes (Decety & Hodges, 2006; Decety and Jackson, 2004; Singer, 2006). Along these lines, meta-analyses reveal that picture-based paradigms elicit activation in the anterior inferior parietal cortex (supramarginal gyrus and intraparietal sulcus) and ventral premotor areas (inferior frontal gyrus, pars opercularis; Figure 12.2, Panel 2; Lamm et al., 2011). Importantly, the joint activation of these two areas is also common to research on action observation (Van Overwalle & Baetens, 2009) and in the hearing or reading of sentences describing action (Aglioti & Pazzaglia, 2010). In fact, sequences of abstract non-biological stimuli (Schubotz & von Cramon, 2004) also activate this network, leading some to propose that it is involved in the prediction of external events (Schubotz, 2007). Given this account, the images in picture-based paradigms may set into motion a cascade of computations that ultimately provide predictive models of affective experience (Lamm et al., 2011). Participants attend to the picture of a knife pressing against a finger, these networks model the knife slicing through the skin that, in turn, elicits affective representations of the consequent laceration. Importantly, this cascade begins with the image of a body part and the implication of an event.

In the absence of images of the body, such a cascade of events is impossible. Instead, vicarious pain necessarily relies upon imagining the state of the other. Accordingly, cue-based vs. image-based events recruit more areas such as the precuneus, ventral parts of the medial prefrontal cortex, the posterior superior temporal cortex, the temporo-parietal junction, and the temporal poles (Figure 12.2, Panel 3; Lamm et al., 2011). These areas are traditionally implicated in Theory of Mind or mentalizing (Gallagher & Frith, 2003; Van Overwalle & Baetens, 2009). It is likely, then, that empathy in cue-based paradigms depends upon mentalizing processes. Participants imagine the condition of the other person and those processes, in turn, elicit empathic states.

The evidence that cue-based and picture-based paradigms elicit empathy via two different routes underscores the fact that understanding others relies on the activation of multiple different networks subserving social cognition (Singer, 2006). In everyday life, the brain uses bodily cues, symbols and pure imaginative inference to elicit empathic experience. While we consider empathic

states as distinctly affective in nature, the mechanisms that lead to empathic responses can rely on computations that would traditionally be labeled as cognitive. This amalgamation of processes allows us to feel empathy in both strictly symbolic circumstances (i.e. while reading a book) and more obviously visceral ones (i.e. while watching a boxing match). Moreover, the fact that various processes elicit empathy suggests that modulating empathic responses, learning empathic skills, and transforming empathy into prosocial behavior may involve a variety of different strategies that tap specific mechanisms.

Distinctions between direct and vicarious pain

One key component in our definition of empathy is that empathizing individuals share the affective state of a target, but preserve the distinction between self and other. The empathizer still identifies the target as the source of the experience. As such, one would expect differences between the neural representations of direct and vicarious pain. At the experiential level this point is obvious; watching a needle puncture someone else's skin can be distressing, but it's not the same feeling as getting pricked yourself. It's not surprising, then, that paradigms designed to directly compare self and vicarious pain find a host of activity that is unique to the direct experience of pain (Lamm et al., 2011).

In particular, cue-based paradigms (e.g. Singer et al., 2004) elicit strong activations in contralateral S1, posterior insula, and contralateral S2 during self pain, but no significant activity in these areas during the vicarious experience of pain. These findings again suggest that the sensory discriminative components of the pain matrix are not necessary for empathic experience. Although picture-based paradigms find higher activity in S1 and S2, those patterns often emerge ipsilaterally as well and on trials in which participants are exposed to both painful and non-painful stimuli (Lamm et al., 2011). Moreover, even patients with a congenital insensitivity to pain display significant activity in bilateral S1 when seeing pictures of others in pain (Danziger, Faille, & Peyron, 2009). Together these data argue against a specific mapping of the somatosensory features during vicarious pain. It is more likely that the activation in somatosensory cortices found in picture-based studies is a consequence of a more general activation elicited by the observation of touch on body parts (Keysers, Kaas, & Gazzola, 2010; Lamm et al., 2011).

Representations of direct and vicarious pain also appear to differ within the insula and cingulate cortex. While both activate anterior portions of the insula, direct pain uniquely activates its posterior subdivisions, which are associated more with sensory features of pain (Figure 12.2, Panel 4; Decety & Lamm, 2009; Lamm et al., 2011). Similarly, direct pain activates a larger portion of the cingulate cortex (Lamm et al., 2011) with distinct activation patterns (Decety & Lamm, 2009; Morrison & Downing, 2007). Connectivity analyses furthermore suggest that overlapping regions responsible for both self and vicarious affect are embedded in larger and divergent networks (Jabbi, Bastiaansen, & Keysers, 2008; Zaki, Ochsner, Hanelin, Wager, & Mackey, 2007).

While it is clear that activation patterns are distinct for direct and vicarious pain, their significant overlap in areas critical for affective experience supports the claim that they rely upon some of the same computations. The spatial resolution of fMRI, however, prevents us from determining whether or not the two states activate the same subpopulations of neurons within the overlapping voxels. Except for one subject in one single cell recording study (Hutchison, Davis, Lozano, Tasker, & Dostrovsky, 1999), there is no direct evidence for precise neuronal overlap between direct and vicarious pain. Nevertheless, recent work using multivariate pattern analysis of fMRI data provides the strongest evidence yet (Corradi-Dell'Acqua et al., 2011). This study looked at multivoxel patterns of activity during direct thermal pain to the hand and the observation of hands in painful situations. Whole brain analyses revealed similar patterns in the AI (bilaterally) in the

two conditions. Region of interest analyses, furthermore, found overlap in the middle insula and middle cingulate cortex. The fact that distributed ensembles of voxels and not simply isolated voxels showed common patterns provides powerful evidence for the shared network hypothesis of empathy for pain.

The role of the anterior insula and cingulate cortex in empathy

The core network found in empathy for pain (the AI and dACC/aMCC) also emerges in research on empathy for other forms of affect. For example, participants in a study on disgust (Jabbi et al., 2008) either tasted a bitter liquid, watched videos of actors tasting bitter liquids, or imagined doing so. All three scenarios elicited activity in the AI and adjacent frontal operculum. Similarly, both inhaling disgusting odorants and seeing faces expressing disgust activated portions of the ACC and AI (Wicker et al., 2003). These regions further emerged in studies on empathic responses to bodies expressing fear (Gelder, Snyder, Greve, Gerard, & Hadjikhani, 2004) and anger (Grosbras & Paus, 2005). Even the sweat of anxious individuals triggered activity in these areas (Pregn-Kristensen, Wiesner, Bergmann, Wolff, & Jansen, 2009). Evidence also suggests that these areas are involved in representing vicarious responses to more obviously social experiences. Specifically, they emerged when participants were exposed to scenes in which targets were embarrassed (Krach, Cohrs, Loebell, Kircher, Sommer, Jansen, et al., 2011) or socially excluded (Masten, Eisenberger, Borofsky, Pfeifer, McNealy, Mazziotta, et al., 2009). Together these data suggest that the AI and dACC/aMCC comprise a network for a multitude of empathic experiences (Bernhardt & Singer, 2012). Given that, what computations occur in these areas and how do they work together to produce empathic experience?

The anterior insula

The insula has long been associated with interoception (Craig, 2002). Functional neuroimaging studies demonstrate its involvement in a wide variety of visceral representations including thirst, bladder distension, sexual arousal, temperature perception, disgust, autonomic arousal, and (of course) pain (Craig, 2009). The AI, specifically, is implicated in the conscious perception of internal states (Craig, 2009), its engagement correlating with interoceptive abilities (Critchley, 2005). For example, AI activity predicts accuracy on a heartbeat detection task in which one compares one's own heartbeat to external feedback. Both performance on this task and self-reports of visceral awareness also correlate with the cortical thickness of the AI (Critchley, Wiens, Rotshtein, Ohman, & Dolan, 2004).

Damasio famously linked the insula's bodily associations with emotional experience. According to his influential model (Damasio, 1994), this region integrates visceral and sensory signals and, in doing so, produces emotional experience. The link between interoceptive awareness and emotional experience has, indeed, been supported by empirical data (Barrett, Quigley, Bliss-Moreau, & Aronson, 2004; Pollatos, Gramann, & Schandry, 2007). Craig (2002, 2009) has further proposed that bodily states are initially represented in the posterior or mid insula, and are then remapped in the AI where they contribute to consciously accessible feeling states. In the domain of pain, these claims are further supported by recent work using direct electrical stimulation of the cortical surface during presurgery evaluations of patients with epilepsy (Mazzola, Isnard, Peyron, & Mauguier, 2011). In over 4000 cortical stimulations of 164 patients, only stimulation of the posterior insula and medial parietal operculum elicited pain responses. Connectivity patterns within the insula and between the insula and other structures further support a posterior-to-anterior mapping of visceral input to conscious and affective remapping (Craig, 2002, 2009).

With its dense connections to both limbic and forebrain regions (Craig, 2009; Critchley et al., 2004; Kurth, Zilles, Fox, Laird, & Eickhoff, 2010), the AI is ideally situated to be a conduit between bodily states and more conscious emotional experience. Accordingly, the AI emerges in multiple studies in which participants focus on their feelings. For example, AI emerged when participants attended to joyful voices (Johnstone, Reekum, Oakes, & Davidson, 2006), or read a sentence expressing a joyful feeling and imagined themselves feeling that joy (Takahashi, Matsuura, Koeda, Yahata, Suhara, Kato, et al., 2008). The role of the AI in affective experience is nicely illustrated by research on alexithymia, a subclinical phenomenon in which individuals have difficulty identifying and describing their emotions. In one study, participants completed a task in which they were exposed to a series of images. Their task on each trial was to either rate their emotional reaction to the image (to introspect) or to simply judge the color balance of the image. Alexithymics showed relatively reduced AI activity when introspecting about their emotional responses to unpleasant images (Silani, Bird, Brindley, Singer, Frith, & Frith, 2008). Similarly, alexithymics showed reduced empathic responses in anterior insula when perceiving close others in pain (Bird, Silani, Brindley, White, Frith, & Singer, 2010).

The AI is probably involved not only in the conscious representation of affective states, but also in computations of prediction and prediction error (Paulus & Stein, 2006; Singer, Critchley, & Preuschoff, 2009). In one study on the anticipation of pain participants completed a series of trials in which they either received painful or non-painful stimulation (Ploghaus, Tracey, Gati, Clare, Menon, Matthews, et al., 1999). Before each trial, the type of stimulation was revealed via a colored light. Participants showed significant AI activity when they saw the pain cue, but before the delivery of the pain, indicating a representation of anticipation of the painful shock. Based on these and other data, researchers have proposed that AI computes predication error between anticipated states and actual visceral input (Paulus & Stein, 2006; Singer et al., 2009). These affective predictions have two critical consequences (Singer et al., 2004; Singer et al., 2009). First, they allow us to anticipate our physiological reactions to emotional stimuli. Secondly, they simulate the affective reactions of other people (i.e. vicarious pain).

Neuroeconomics research further implicates the AI in processing and prediction of risk and uncertainty (Critchley, Mathias, & Dolan, 2001; Grinband, Hirsch, & Ferrera, 2006; Paulus, Rogalsky, Simmons, Feinstein, & Stein, 2003; Preuschoff, Quartz, & Bossaerts, 2008). The AI is active during tasks which are risky, ambiguous, or complex (Elliott, Friston, & Dolan, 2000; Grinband et al., 2006; Huettel, Stowe, Gordon, Warner, & Platt, 2006). These data suggest that the AI predicts risk and uncertainty and computes errors between those predictions and actual outcomes (Singer et al., 2009). For example, bilateral AI activity emerged when participants waited for the outcome of a risky decision and the level of activity reflected the risk prediction error once the outcome was known (Preuschoff et al., 2008).

Based on these findings and on the involvement of the AI in representing direct and vicarious feeling states, Singer and colleagues have suggested a broader model of AI functioning (Singer et al., 2009). Within this model the AI integrates information about online and projected feeling states. It processes incoming sensory, bodily, and contextual information, while generating predictions for the affective consequences of anticipated events. By comparing these two channels of data, it calculates and refines estimates of outcomes, uncertainty, and their prediction errors. Together these functions produce a global feeling state, which reflects the integration of interoception, prediction, and risk. Critically, this integration would allow the AI to drive emotional learning and decision making. When considered in terms of empathy, the AI may compute the projected feelings states of another person and may, furthermore, compare those states with online information (e.g. facial or vocal expressions, bodily state, and etc.). Such computations would allow us to learn from others'

positive or negative experiences and to, perhaps, provide help or support when they are needed. In other words, we can learn and make decisions, not only from our own emotional states, but from the observed or imagined states of others.

The cingulate cortex

The cingulate cortex also emerges across studies on empathy, specifically the dACC/aMCC. Functionally, this region has been implicated in a wide variety of phenomena. A recent meta-analysis of 939 studies found that overlapping portions of the dACC/aMCC are involved in representing negative affect, pain, and cognitive control (Shackman, Salomons, Slagter, Fox, Winter, & Davidson, 2011). Other data also implicate the regions in response selection (Medford & Critchley, 2010). Researchers consistently find concurrent activation in the AI and these regions of the cingulate, particularly in emotion-related paradigms (Craig, 2009; Medford & Critchley, 2010). In line with those findings, resting state fMRI connectivity analyses show a close functional relationship between the AI and these areas (Taylor, Seminowicz, & Davis, 2009; Harrison, Pujol, Ortiz, Fornito, Pantelis, & Yucel, 2008), a relationship that is supported by dense anatomical interconnections (Bernhardt & Singer, 2012; Medford & Critchley, 2010).

As mentioned above, these portions of the ACC contribute to the affective component of the pain matrix (Apkarian et al., 2005; Rainville et al., 1997). Specifically, they're associated with the motivational component of the response and likely play a critical role in preparing responsive action (Morrison & Downing, 2007; Vogt & Sikes, 2009). In line with this account, the dACC/aMCC receives direct projections from pain pathways; caudal divisions of the cingulate near the dACC/aMCC also have strong functional connections to sensorimotor regions (Margulies, Kelly, Uddin, Biswal, Castellanos, & Milham, 2007). Research on animal models provides evidence that these regions are involved in motivated action. Ablation of the ACC in rats selectively reduces avoidant behavior without reducing sensitivity to noxious stimuli (LaGraize Labuda, Rutledge, Jackson, & Fuchs, 2004). Furthermore, single neuron recordings in monkeys have identified neurons in the ACC that selectively fire in response to cues for forthcoming pain or reward stimuli that they can either approach or avoid (Koyama, Keichiro, Tanaka, & Mikami, 2001).

The joint activation of AI and regions of the ACC also emerges frequently in studies on emotion (Craig, 2009; Medford & Critchley, 2010). As with pain, dACC/aMCC activity here may represent the mapping of affective responses into motivational and somatic domains (Craig, 2009; Pollatos et al., 2007; Medford & Critchley, 2010). Accordingly, joint ACC and AI activity has been shown in response to emotional facial expressions and, more to the point, is associated with heart rate changes (Critchley, Rotshtein, Nagai, O'Doherty, Mathias, & Dolan, 2005a). Further evidence links aMCC activity more generally with autonomic arousal (Critchley, Tang, Glaser, Butterworth, & Dolan, 2005b). Within the domain of empathy, the link between the AI and the dACC/aMCC likely represents the causal chain from projections of the target's feeling state to motivational and behavioral responses. In other words, feeling that others are in distress can drive us to flinch, cringe, or act. Of course the nature of the behavioral response varies. We will return to the behavioral consequences of empathy at the end of this chapter.

Empathy in somatosensation

Although we have focused on the role of the AI and ACC in empathy, research also demonstrates empathy for touch in regions more directly associated with somatosensation (Blakemore, Bristow, Bird, Frith, & Ward, 2005; Ebisch, Perrucci, Ferretti, Gratta, Romani, & Gallese, 2008; Keysers et al., 2004, 2010; see also Keysers Thioux, and Gazzola, this volume). Participants have consistently shown common activation in the secondary somatosensory cortex when they both experience

and observe touch. In one fMRI study, participants watched videos of legs being stroked, or had their own legs stroked in a similar fashion. This overlap in activation within in the secondary somatosensory cortex in the two conditions (Keysers et al., 2004) has been replicated in replicated in subsequent research (Ebisch et al., 2008; Schaefer, Xu, Flor, & Cohen, 2009). One study, furthermore, identified an individual who experiences conscious tactile sensation when watching another person being touched (Ebisch et al., 2008). While both she and normal controls showed responses to observed touch in the somatosensory cortices, the activity was significantly greater for this individual. Together these data demonstrate that the observation and experience of somatosensation recruits common networks.

The modulation of empathic responses

It is doubtful that anyone feels empathy for all people at all times. In fact, it is quite easy to come up with situations in which we feel more or less empathy toward an individual based on who that person is or how they have behaved. There are also clear differences between individuals in the ability or motivation to empathize with others. Given this wide variability, understanding the modulation of empathy is critical to understanding the phenomenon itself. Along these lines, researchers have explored the effects of context, interpersonal factors, and individual empathic capacities to better understand how and why empathic responses vary (see also de Vignemont and Singer, 2006; Hein & Singer, 2008).

Our ability and willingness to empathize with others is strongly affected by their identity and behavior. While the original cue-based, empathy-for-pain paradigm described above looked at empathy for loved ones (Singer et al., 2004), follow-up research has used the same paradigm to examine empathy for strangers. In one such study (Singer et al., 2006), participants came into the lab and interacted with confederates whom they believed to be other participants and who differed in their fair or unfair behavior toward the participant. They first completed an economic game with the confederates. During the game, one confederate played fairly and the other participant played unfairly, “defecting” in economic exchanges by failing to reciprocate the participant’s offers. Later, the participant and confederates completed the empathy-for-pain task. When fair players received painful shocks, participants exhibited the same empathic response exhibited in the prior study (i.e. activity in the AI and ACC). When unfair players received shocks, however, male participants showed relatively reduced responses in these regions. Moreover, they exhibited increased activity in areas that have typically been associated with reward processing (i.e. the nucleus accumbens). These increases also correlated with the self-reported desire for revenge.

This effect was replicated in a subsequent study in which participants witnessed both ingroup and outgroup members receiving a painful stimulus (Hein, Silani, Preuschoff, Batson, & Singer, 2010). Participants were soccer fans and they interacted with fellow fans of their favorite team (in-group members) and fans of their favorite team’s rival (out-group members). The results showed stronger responses in the left AI when participants witnessed an in-group member vs. an out-group member suffer. As with unfair players in the earlier study, witnessing out-group members elicited activation in the nucleus accumbens that was modulated by group perception.

Characteristics of a target person’s perceived affective state can also moderate the empathic response. In one study (Saarela et al., 2007), participants were shown photos of faces of chronic pain patients who were experiencing varying levels of acute pain. Participants showed more activity in several areas including the AI and ACC when exposed to the acute pain photos. Moreover, their estimates of targets’ pain intensity correlated with the strength of activation in these areas (left ACC, left inferior parietal lobule (IPL), and bilateral AI). Here, the affective facial cues modulated the empathic responses, even in the absence of bodily cues about the painful stimulus itself.

Conversely, contextual information can alter empathic responses to identical images of the body. In one study, participants were shown similar images of hands undergoing medical procedures (Lamm et al., 2007b). One image type depicted a painless biopsy performed on an anesthetized hand while the other depicted a painful injection into a hand. Despite the relatively abstract information about the nature of the photos, participants showed reduced empathic responses in the AI and aMCC when exposed to the anesthetized vs. non-anesthetized hand. Knowing the consequences of a painful event can also affect the empathic neural response. For example, when participants watched videos of a painful procedure, they showed weaker activity in the aMCC and AI when they believed that procedure to be therapeutically effective than when they believed it to be ineffective.

Attention and imagination also play critical parts in the modulation of empathy. Participants who observed images of hands in painful situations showed stronger activation in the AI and ACC when they focused on the intensity of the person's pain as opposed to physical features of the image (Gu & Han, 2007). Similarly, perspective-taking can alter the neural response. Participants imagining themselves in a painful situation vs. imagining another person in that situation show enhanced responses to the images, notably in the insula and aMCC (Jackson et al., 2006; Lamm et al., 2007a).

As mentioned in the previous section, individual differences in self-report measures of empathy (Baron-Cohen & Wheelwright, 2004; Davis, 1983; Mehrabian & Epstein, 1972) correlate with empathic responses to equivalent stimuli. Conversely, empathic deficits, on the other hand, emerge in various clinical phenomena. Of course the most obvious disorder for which we would expect deficits is psychopathy. While there is not yet direct evidence for a reduced empathic neural response in psychopaths, less AI grey matter volume has been associated with weaker empathy scores in adolescents with conduct disorder (Sterzer, Stadler, Poustka, & Kleinschmidt, 2007). Research on adult psychopaths, furthermore, has shown reduced activity of the amygdala and AI during the anticipation of pain (Birbaumer, Veit, Lotze, Erb, & Hermann, 2005). More data are necessary to make a claim about the neural nature of empathic deficits in psychopathy.

The data on empathic deficits in alexithymics are clearer. As mentioned above, alexithymics have reduced introspective abilities which appear to translate into reduced empathic responses (Silani et al., 2008). Silani and colleagues' findings have been replicated in subsequent research (Bird et al., 2010). Again, empathic neural activation elicited by the pain of a close other was modulated by individual levels of alexithymia. Importantly, this sample included individuals with autism spectrum conditions. When analyses accounted for levels of alexithymia, empathic responses were comparable between autistic and control groups. As such, although alexithymia and autism spectrum disorders show high comorbidity, there is no necessary deficit in empathy in autism. This double dissociation further underscores the distinction between empathic and mentalizing abilities as autism spectrum disorders are known to be associated with severe theory of mind deficits (Baron-Cohen, 1995).

While we commonly consider empathy to be a positive trait, there are some domains in which a controlled empathic response is clearly beneficial. Health practitioners, for example, would have a terrible time if they winced or cringed every time they had to perform a painful procedure. One study by Cheng and colleagues (2007) addressed this point, exposing both acupuncturists and laymen to images of needles being inserted into different parts of the body. As predicted, only the laymen showed neural activation characteristic of empathic responses. Of course, it seems likely that while acupuncturists may control their empathic responses to pain, they likely preserve empathic responses in other domains. After all, different circumstances require different responses. Along those lines, the complex relationship between empathy and behavior is the topic of our final section.

From empathy to prosocial behavior

Although our definition of empathy does not refer directly to behavior, one would expect such a basic component of social life to influence it. Indeed, the network of regions we've focused on here suggest a causal pathway from other-oriented prediction to conscious feeling state to motivation (Bernhardt & Singer, 2012; Craig, 2009). As such, empathy likely prepares one to respond and, possibly, to act. At the individualistic level, sharing other peoples' feelings allows us react to their distress or joy so that we can avoid their mistakes or emulate their successes. In more prosocial terms, empathy allows us to respond to the needs of distressed others or to share in their joy.

These two putative functions of empathy can imply very different behavioral consequences. Accordingly, empathy researchers have long drawn a key distinction between two different empathic reactions: empathic concern and empathic distress (Batson, 2009a; Eisenberg, 2000; Klimecki & Singer, 2012). Empathic concern is akin to sympathy. The concerned individual responds to the distressed state of another, but with an approach motivation—they feel a desire to care for the target. Empathic distress, on the other hand, is an aversive state associated with avoidance motivation. The empathically distressed individual assumes the distressed feelings of the target to such an extent that they must physically or symbolically flee the situation. They are incapable of helping.

Social neuroscience is only beginning to explore this difference and to better understand how empathy might lead to the kind of approach behaviors associated with helping. Lamm and colleagues (2007a), for example, point to the fact that when participants consider a painful scene using self (vs. other) perspective-taking, they show a stronger activation in components of the core empathy network (the insula and the aMCC) and in the amygdala (which among other things plays a critical role in fear-related behaviors; LeDoux, 2003). The assumption of the first person perspective here may push the experience into empathic distress such that the individual experiences the kinds of personal distress and avoidance motivation associated with direct pain.

On the other hand, Hein and colleagues (2010) have provided neural evidence that empathy can motivate costly helping. As part of the abovementioned study on empathy for in-group vs. out-group members, participants had the opportunity to receive a painful stimulus in order to reduce the painful stimulus delivered to another player. Participants who showed more AI activation while seeing an in-group member suffer were more likely to help that person. Conversely, participants who showed more nucleus accumbens activity (associated with reward), while seeing the out-group member suffer, were less likely to help.

Work on social exclusion extends these findings (Masten, Morelli, & Eisenberger, 2010). During this fMRI experiment, participants observed one person being excluded by two other people during a computerized ball-tossing game (Williams et al., 2000). After the scanning period, participants were asked to send emails to the players whom they had observed. Coders rated the degree to which the emails sent to the ostracized individual were comforting, supportive, and attempted to be helpful. Analyses revealed positive relationships between these prosocial communications and activity in the right AI and the medial prefrontal cortex during the exclusionary event (Masten et al., 2010; see also Mathur, Harada, Lipke, & Chiao, 2010). Given these data, it appears that empathic experience, and its neural components, can indeed promote helping behaviors.

Outlook

A crucial question for the future study of empathy is how and why vicarious feelings sometimes lead to empathic distress and avoidance, and other times lead to empathic concern and helping. Emotion regulation likely plays an important role in these processes, but which specific regulatory

strategies are effective, the nature of their neural representations, and how they interact with neural representations of empathy remain to be seen.

Important questions also remain regarding the relationship between empathy and different psychopathologies. For example, how does empathy manifest (or fail to manifest) in individuals with depression, borderline personality disorder, or narcissistic personality disorder (see Ritter, Dziobek, Preissler, Rütter, Vater, & Fydrich, 2011)? It will also be critical to disentangle ways in which mentalizing pathways and empathy pathways are differentially affected in these disorders. Besides the obvious application in clinical domains, this line of research will help distinguish between the various mechanisms that drive social cognition.

The plasticity of empathy is another key frontier. Can one be trained to be more empathic or to better transform the empathic response into prosocial action? If so, what are the components of effective empathy training? Along similar lines, neuroscience is only beginning to investigate the developmental trajectory of empathy (e.g. Decety, Michalska, & Akitsuki, 2008) and, more generally, social emotions. Future work in this domain will help identify the critical periods in which social emotions emerge, and the factors that facilitate their emergence.

Ideally, these new lines of inquiry will translate the basic findings from the neuroscience of empathy into everyday benefits. Empathy, after all, is one of human nature's more appealing traits.

Acknowledgements

We would like to thank Ralf Hartmann for his assistance in creating Figure 12.1. We also thank Claus Lamm and Boris Bernhardt for their help with Figure 12.2.

References

- Aglioti, S.M., & Pazzaglia, M. (2010). Representing actions through their sound. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale* 206: 141–51.
- Apkarian, A.V., Bushnell, M.C., Treede, R.D., & Zubieta, J.K. (2005). Human brain mechanisms of pain perception and regulation in health and disease. *European Journal of Pain* 9: 463–84.
- Avenanti, A., Buetti, D., Galati, G., & Aglioti, S.M. (2005). Transcranial magnetic stimulation highlights the sensorimotor side of empathy for pain. *Nature neuroscience* 8: 955–60.
- Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge: MIT Press/Bradford Books.
- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism Development Disorders* 34: 163–75.
- Barrett, L.F., Quigley, K.S., Bliss-Moreau, E., & Aronson, K. R. (2004). Interoceptive sensitivity and self-reports of emotional experience. *Journal of Personal Social Psychology* 87: 684–97.
- Batson, C. D. (2009a). Empathy-induced altruistic motivation. In: M. Mikulincer & P. R. Shaver (Eds), *Prosocial Motives, Emotions, and Behavior* (pp. 15–34). Washington, D.C.: American Psychological Association.
- Batson, C. D. (2009b). These things called empathy: eight related, but distinct phenomena. In: J. Decety & W. Ickes (Eds), *The Social Neuroscience of Empathy* (pp. 3–15). Cambridge: MIT Press.
- Batson, C.D., Fultz, J. & Schoenrade, P.A. (1987). Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of Personality* 55: 19–39.
- Baumeister, R.F. & Vohs, K.D. (2007). *Encyclopedia of Social Psychology*. Thousand Oaks: Sage Publications.
- Bernhardt, B. & Singer, T. (2012). The neural basis of empathy. *Annual Reviews of Neuroscience* 35: 1–23.

- Bingel, U., Quante, M., Knab, R., Bromm, B., Weiller, C. & Buchel, C. (2003). Single trial fMRI reveals significant contralateral bias in responses to laser pain within thalamus and somatosensory cortices. *NeuroImage* 18: 740–8.
- Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., & Flor, H. (2005). Deficient fear conditioning in psychopathy: a functional magnetic resonance imaging study. *Archives of General Psychiatry* 62: 799–805.
- Bird, G., Silani, G., Brindley, R., White, S., Frith, U. & Singer, T. (2010). Empathic brain responses in insula are modulated by levels of alexithymia, but not autism. *Brain: Journal of Neurology* 133: 1515–25.
- Blakemore, S.J., Bristow, D., Bird, G., Frith, C., & Ward, J. (2005). Somatosensory activations during the observation of touch and a case of vision-touch synaesthesia. *Brain* 128: 1571–83.
- Botvinick, M., Jha, A.P., Bylsma, L.M., Fabian, S.A., Solomon, P.E., & Prkachin, K.M. (2005). Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. *NeuroImage* 25: 312–19.
- Brooks, J.C., Nurmikko, T.J., Bimson, W.E., Singh, K.D., & Roberts, N. (2002). fMRI of thermal pain: effects of stimulus laterality and attention. *NeuroImage* 15: 293–301.
- Buccino, G., Binkofski, F., Fink, G.R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R.J., Zilles, K., Rizzolatti, G. & Freund, H.J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience* 13: 400–4.
- Cheng, Y., Lin, C.P., Liu, H.L., Hsu, Y.Y., Lim, K.E., Hung, D., & Decety, J. (2007). Expertise modulates the perception of pain in others. *Current Biology* 17: 1708–13.
- Corradi-Dell'Acqua, C., Hofstetter, C., & Vuilleumier, P. (2011). Felt and seen pain evoke the same local patterns of cortical activity in insular and cingulate cortex. *Journal of Neuroscience* 31: 17966–8006.
- Craig, A.D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *National Review of Neuroscience* 3: 655–66.
- Craig, A.D. (2009). How do you feel—now? The anterior insula and human awareness. *National Review of Neuroscience* 10: 59–70.
- Critchley, H.D. (2005). Neural mechanisms of autonomic, affective, and cognitive integration. *Journal of Comprehensive Neurology* 493: 154–66.
- Critchley, H.D., Mathias, C.J., & Dolan, R.J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron* 29: 537–45.
- Critchley, H.D., Rotshtein, P., Nagai, Y., O'Doherty, J., Mathias, C.J., & Dolan, R.J. (2005a). Activity in the human brain predicting differential heart rate responses to emotional facial expressions. *NeuroImage* 24: 751–62.
- Critchley, H.D., Tang, J., Glaser, D., Butterworth, B., & Dolan, R.J. (2005b). Anterior cingulate activity during error and autonomic response. *NeuroImage* 27: 885–95.
- Critchley, H.D., Wiens, S., Rotshtein, P., Ohman, A. & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience* 7: 189–95.
- Damasio, A.R. (1994). Descartes' error and the future of human life. *Scientific American* 271: 144.
- Danziger, N., Faillenot, I., & Peyron, R. (2009). Can we share a pain we never felt? Neural correlates of empathy in patients with congenital insensitivity to pain. *Neuron* 61, 203–12.
- Davis, M.H. (1983). Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of Personality and Social Psychology* 44: 113–26.
- de Vignemont, F., & Singer, T. (2006). The empathic brain: how, when and why? *Trends in Cognitive Sciences* 10: 435–41.
- Decety, J., & Hodges, S.D. (2006). The social neuroscience of empathy. In: P. A. M. Lange (Ed.), *Bridging Social Psychology Benefits of Transdisciplinary Approaches* (pp. 103–10). Mahwah: Erlbaum.
- Decety, J. & Jackson, P.L. (2004). The functional architecture of human empathy. *Behavioral Cognitive Neuroscience Review* 3: 71–100.

- Decety, J. & Lamm, C. (2009). Empathy and intersubjectivity. In J. T. Cacioppo & G. G. Bernston (Eds), *Handbook of Neuroscience for the Behavioral Sciences* (pp. 940–57). New York: John Wiley and Sons.
- Decety, J., Michalska, K. & Akitsuki, Y. (2008). Who caused the pain? An fMRI investigation of empathy and intentionality in children. *Neuropsychologia* 46: 2607–14.
- Derbyshire, S.W. (2000). Exploring the pain “neuromatrix”. *Current Review of Pain* 4: 467–77.
- di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research* 91: 176–80.
- Doherty, R. (1997). The emotional contagion scale: A measure of individual differences. *Journal of Non-verbal Behavior* 21: 123.
- Ebisch, S., Perrucci, M., Ferretti, A., Gratta, C.D., Romani, G., & Gallese, V. (2008). The sense of touch: Embodied simulation in a visuotactile mirroring mechanism for observed animate or inanimate touch. *Journal of Cognitive Neuroscience* 20: 12.
- Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annual Review of Psychology* 51: 665–97.
- Eisenberg, N., & Fabes, R.A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion* 14: 131–49.
- Elliott, R., Friston, K.J., & Dolan, R.J. (2000). Dissociable neural responses in human reward systems. *Journal of Neuroscience* 20: 6159–65.
- Fan, Y., Duncan, N., Greck, M.D., & Northoff, G. (2011). Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience and Biobehavioral Reviews* 35: 11.
- Gallagher, H.L., & Frith, C.D. (2003). Functional imaging of ‘theory of mind’. *Trends in Cognitive Science* 7: 77–83.
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119(Pt 2): 593–609.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Science* 2: 493–501.
- Gelder, B.D., Snyder, J., Greve, D., Gerard, G., & Hadjikhani, N. (2004). Fear fosters flight: a mechanism for fear contagion when perceiving emotion expressed by a whole body. *Proceedings of the National Academy of Science USA* 101: 6.
- Grinband, J., Hirsch, J., & Ferrera, V.P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron* 49: 757–63.
- Grosbras, M.H., & Paus, T. (2005). Brain networks involved in viewing angry hands or faces. *Cerebral Cortex*, 16(8): 1087–96.
- Gu, X., & Han, S. (2007). Attention and reality constraints on the neural processes of empathy for pain. *Neuroimage* 36: 256–67.
- Harrison, B., Pujol, J., Ortiz, H., Fornito, A., Pantelis, C. & Yücel, M. (2008). Modulation of brain resting-state networks by sad mood induction. *PLoS One* 3: e1794.
- Hatfield, E., Cacioppo, J.T., & Rapson, R. (1994). *Emotional Contagion*, New York: Cambridge University Press.
- Hatfield, E., Rapson, R., & Le, Y. (2009). Emotional contagion and empathy. In J. Decety & W. Ickes (Eds), *The Social Neuroscience of Empathy* (pp. 19–30), Boston: MIT Press.
- Hein, G., Silani, G., Preuschoff, K., Batson, C.D., & Singer, T. (2010). Neural responses to ingroup and outgroup members’ suffering predict individual differences in costly helping. *Neuron* 68: 149–60.
- Hein, G., & Singer, T. (2008). I feel how you feel, but not always: the empathic brain and its modulation. *Current Opinion in Neurobiology* 18: 153–8.
- Huettel, S., Stowe, C.J., Gordon, E.M., Warner, B.T. & Platt, M. (2006). Neural signatures of economic preferences for risk and ambiguity. *Neuron* 49: 765–75.
- Hutchison, W.D., Davis, K.D., Lozano, A.M., Tasker, R.R., & Dostrovsky, J.O. (1999). Pain-related neurons in the human cingulate cortex. *Nature Neuroscience* 2: 403–5.

- Jabbi, M., Bastiaansen, J., & Keysers, C. (2008). A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PLoS One* 3: e2939.
- Jabbi, M., Swart, M., & Keysers, C. (2007). Empathy for positive and negative emotions in the gustatory cortex. *Neuroimage* 34: 1744–53.
- Jackson, P.L., Brunet, E., Meltzoff, A.N., & Decety, J. (2006). Empathy examined through the neural mechanisms involved in imagining how I feel vs. how you feel pain. *Neuropsychologia* 44: 752–61.
- Jackson, P.L., Meltzoff, A.N., & Decety, J. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. *Neuroimage* 24: 771–9.
- Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage* 14: 103–9.
- Johnstone, T., Reekum, C.V., Oakes, T., & Davidson, R. (2006). The voice of emotion: an fMRI study of neural responses to angry and happy vocal expressions. *Social Cognitive and Affective Neuroscience* 1: 242–9.
- Keysers, C., & Gazzola, V. (2007). Integrating simulation and theory of mind: from self to social cognition. *Trends in Cognitive Science* 11: 194–6.
- Keysers, C., Kaas, J.H., & Gazzola, V. (2010). Somatosensation in social perception. *National Review of Neuroscience* 11: 417–28.
- Keysers, C., Wicker, B., Gazzola, V., Anton, J.L., Fogassi, L., & Gallese, V. (2004). A touching sight: SII/PV activation during the observation and experience of touch. *Neuron* 42: 335–46.
- Klimecki, O., & Singer, T. (2013). Empathy from the perspective of social neuroscience. In J. Armony & P. Vuilleumier (Eds), *Handbook of Human Affective Neuroscience* (pp. 533–51). New York: Cambridge University Press.
- Kong, J., Gollub, R.L., Polich, G., Kirsch, I., Laviolette, P., Vangel, M., Rosen, B., & Kaptchuk, T.J. (2008). A functional magnetic resonance imaging study on the neural mechanisms of hyperalgesic placebo effect. *Journal of Neuroscience* 28: 13354–62.
- Koyama, T., Keichiro, K., Tanaka, Y., & Mikami, A. (2001). Anterior cingulate activity during pain-avoidance and reward in monkeys. *Neuroscience Research* 39: 421–30.
- Krach, S., Cohrs, J., Loebell, N.D.E., Kircher, T., Sommer, J., Jansen, A., Paulus, F.M. (2011). Your flaws are my pain: linking empathy to vicarious embarrassment. *PLoS One* 6: e18675.
- Kurth, F., Zilles, K., Fox, P.T., Laird, A.R., & Eickhoff, S.B. (2010). A link between the systems: functional differentiation and integration within the human insula revealed by meta-analysis. *Brain Structure and Function* 214: 519–534.
- Lagraize, S., Labuda, C., Rutledge, M., Jackson, R., & Fuchs, P. (2004). Differential effect of anterior cingulate cortex lesion on mechanical hypersensitivity and escape/avoidance behavior in an animal model of neuropathic pain. *Experimental Neurology* 188: 139–48.
- Lamm, C., Batson, C.D., & Decety, J. (2007a). The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience* 19: 42–58.
- Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage* 54: 2492–502.
- Lamm, C., Meltzoff, A.N., & Decety, J. (2010). How do we empathize with someone who is not like us? A functional magnetic resonance imaging study. *Journal of cognitive neuroscience* 22: 362–76.
- Lamm, C., Nusbaum, H.C., Meltzoff, A.N., & Decety, J. (2007b). What are you feeling? Using functional magnetic resonance imaging to assess the modulation of sensory and affective responses during empathy for pain. *PLoS one* 2: e1292.
- Ledoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology* 23: 727–38.
- Maihöfner, C., Herzner, B., & Handwerker, H. (2006). Secondary somatosensory cortex is important for the sensory-discriminative dimension of pain: A functional MRI study. *European Journal of Neuroscience* 23: 1377–83.

- Margulies, D., Kelly, A., Uddin, L., Biswal, B., Castellanos, F., & Milham, M. (2007). Mapping the functional connectivity of anterior cingulate cortex. *NeuroImage* 37: 579–88.
- Masten, C., Morelli, S., & Eisenberger, N. (2010). An fMRI investigation of empathy for 'social pain' and subsequent prosocial behavior. *NeuroImage* 55, 381–8.
- Masten, C.L., Eisenberger, N.I., Borofsky, L.A., Pfeifer, J.H., McNealy, K., Mazziotta, J.C. & Dapretto, M. (2009). Neural correlates of social exclusion during adolescence: understanding the distress of peer rejection. *Social Cognitive and Affective Neuroscience* 4: 143–57.
- Mathur, V., Harada, T., Lipke, T. & Chiao, J. (2010). Neural basis of extraordinary empathy and altruistic motivation. *NeuroImage* 51: 1468–75.
- Mazzola, L., Isnard, J., Peyron, R. & Mauguiere, F. (2011). Stimulation of the human cortex and the experience of pain: Wilder Penfield's observations revisited. *Brain* 135(Pt 2): 631–40.
- Medford, N., & Critchley, H.D. (2010). Conjoint activity of anterior insular and anterior cingulate cortex: awareness and response. *Brain Structure & Function* 214: 535–49.
- Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality* 40: 525–43.
- Melzack, R. (1975). The McGill Pain Questionnaire: Major properties and scoring methods. *Pain* 1: 22.
- Morrison, I. & Downing, P.E. (2007). Organization of felt and seen pain responses in anterior cingulate cortex. *NeuroImage* 37: 642–51.
- Morrison, I., Lloyd, D., di Pellegrino, G., & Roberts, N. (2004). Vicarious responses to pain in anterior cingulate cortex: Is empathy a multisensory issue? *Cognitive Affective Behavioral Neuroscience* 4: 270–8.
- Paulus, M.P., Rogalsky, C., Simmons, A., Feinstein, J.S., & Stein, M.B. (2003). Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. *NeuroImage* 19: 1439–48.
- Paulus, M.P., & Stein, M.B. (2006). An insular view of anxiety. *Biological Psychiatry* 60: 383–7.
- Peyron, R., Laurent, B., & Garcia-Larrea, L. (2000). Functional imaging of brain responses to pain. A review and meta-analysis (2000). *Neurophysiology Clinic* 30: 263–88.
- Ploghaus, A., Tracey, I., Gati, J.S., Clare, S., Menon, R.S., Matthews, P.M., & Rawlins, J.N. (1999). Dissociating pain from its anticipation in the human brain. *Science* 284: 1979–81.
- Ploner, M., Freund, H., & Schnitzler, A. (1999). Pain affect without pain sensation in a patient with a post-central lesion. *Pain* 81: 211–14.
- Pollatos, O., Gramann, K., & Schandry, R. (2007). Neural systems connecting interoceptive awareness and feelings. *Human Brain Mapping* 28: 9–18.
- Prehn-Kristensen, A., Wiesner, C., Bergmann, T., Wolff, S., & Jansen, O. (2009). Induction of empathy by the smell of anxiety. *PLoS One* 4: e5987.
- Preston, S.D. & de Waal, F.B. (2002). Empathy: Its ultimate and proximate bases. *Behavioural Brain Science* 25: 1–20; discussion 20–71.
- Preuschoff, K., Quartz, S.R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience* 28: 2745–2752.
- Price, D. 2000. Psychological and neural mechanisms of the affective dimension of pain. *Science* 288: 1769–72.
- Prinz, W. (1997). Perception and action planning. *Journal of Cognitive Psychology* 9: 129–54.
- Prinz, W. (2005). Experimental approaches to action. In: J. Roessler & N. Eilan (Eds), *Agency and Self-Awareness* (pp. 165–87). New York: Oxford University Press.
- Rainville, P., Duncan, G.H., Price, D.D., Carrier, B., & Bushnell, M.C. (1997). Pain affect encoded in human anterior cingulate, but not somatosensory cortex. *Science* 277: 968–71.
- Ritter, K., Dziobek, I., Preissler, S., Rütter, R., Vater, A., Fydrich, T., Lammers, C., Heekeren, H., & Roepke, S. (2011). Lack of empathy in patients with narcissistic personality disorder. *Psychiatry Research* 187: 241–7.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience* 27: 169–92.

- Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience* 2: 661–70.
- Saarela, M.V., Hlushchuk, Y., Williams, A.C., Schürmann, M., Kalso, E., & Hari, R. (2007). The compassionate brain: humans detect intensity of pain from another's face. *Cerebral Cortex* 17: 230–7.
- Schaefer, M., Xu, B., Flor, H., & Cohen, L. (2009). Effects of different viewing perspectives on somatosensory activations during observation of touch. *Human Brain Mapping* 30: 2722–30.
- Schubotz, R.I. (2007). Prediction of external events with our motor system: toward a new framework. *Trends in Cognitive Sciences* 11: 211–18.
- Schubotz, R.I., & von Cramon, D.Y. (2004). Sequences of abstract non-biological stimuli share ventral premotor cortex with action observation and imagery. *Journal of Neuroscience* 24: 5467–74.
- Shackman, A., Salomons, T., Slagter, H., Fox, A., Winter, J., & Davidson, R. (2011). The integration of negative affect, pain and cognitive control in the cingulate cortex. *National Review of Neuroscience* 12: 13.
- Silani, G., Bird, G., Brindley, R., Singer, T., Frith, C., & Frith, U. (2008). Levels of emotional awareness and autism: an fMRI study. *Social Neuroscience* 3: 97–112.
- Singer, T. (2006). The neuronal basis and ontogeny of empathy and mind reading: review of literature and implications for future research. *Neuroscience and Biobehavioral Reviews* 30: 855–63.
- Singer, T., Critchley, H.D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences* 13: 334–40.
- Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences* 1156: 81–96.
- 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: 1157–62.
- Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., & Frith, C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439: 466–9.
- Singer, T., Snozzi, R., Bird, G., Petrovic, P., Silani, G., Heinrichs, M., & Dolan, R.J. (2008). Effects of oxytocin and prosocial behavior on brain responses to direct and vicariously experienced pain. *Emotion* 8: 781–91.
- Sterzer, P., Stadler, C., Poustka, F., & Kleinschmidt, A. (2007). A structural neural deficit in adolescents with conduct disorder and its association with lack of empathy. *NeuroImage* 37: 335–42.
- Takahashi, H., Matsuura, M., Koeda, M., Yahata, N., Suhara, T., Kato, M., & Okuba, Y. (2008). Brain activations during judgments of positive self-conscious emotion and positive basic emotion: Pride and joy. *Cerebral Cortex* 18: 893–903.
- Taylor, K., Seminowicz, D., & Davis, K. (2009). Two systems of resting state connectivity between the insula and cingulate cortex. *Human Brain Mapping* 30: 14.
- van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage* 48: 564–84.
- Vogt, B., & Sikes, R. (2009). Cingulate nociceptive circuitry and roles in pain processing: the cingulate premotor pain model. In: B. Vogt (Ed.), *Cingulate Neurobiology and Disease* (pp. 311–38). Oxford: Oxford University Press.
- Wicker, B., Keysers, C., Plailly, J., Royet, J.P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in my insula: the common neural basis of seeing and feeling disgust. *Neuron* 40: 655–64.
- Williams, K., Cheung, C., & Choi, W. (2000). Cyberostracism: effects of being ignored over the Internet. *Journal of Personal and Social Psychology* 79: 748–62.
- Wispe, L. (1986). The distinction between sympathy and empathy: to call forth a concept, a work is needed. *Journal of Personality and Social Psychology* 50: 314–21.
- Zaki, J., Ochsner, K.N., Hanelin, J., Wager, T.D., & Mackey, S.C. (2007). Different circuits for different pain: patterns of functional connectivity reveal distinct networks for processing pain in self and others. *Social Neuroscience* 2: 276–91.

Neural sources of empathy: An evolving story

Jamil Zaki and Kevin Ochsner

Compared with many other animals on the planet, human beings are small, slow, soft, and weak. Yet, we have unequivocally won the cross-species competition for global domination. What allowed us, as physical underdogs, to claim this unlikely victory? In other words, what makes humans special?

While many disciplines have addressed this question in some way, psychologists' answer has evolved over time. Until recently, the dominant view held that human uniqueness was bound up in our *intrapersonal* abilities, such as the use of arbitrary symbols (Deacon, 1997; Pinker, 1994; Pinker & Bloom, 1990), and recursive syntax (Chomsky, 1980; Hauser, Chomsky, & Fitch, 2002) in language, or our ability to mentally "time travel" in reflecting on past experiences and planning future actions (Suddendorf & Corballis, 1997, 2007; Tulving, 2002). Although these faculties are undoubtedly critical, in the last decade human specialness has come to be seen as much more interpersonal: embodied, for example, in our abilities to understand (Brothers, 1997; Leslie, 1994), learn from (Csibra & Gergely, 2006; Moll & Tomasello, 2007), and share intentions with others (Tomasello, 2000; Tomasello, Carpenter, Call, Behne, & Moll, 2005).

Together, such abilities contribute to the multi-faceted construct of empathy. Empathy is thought to comprise multiple related, but distinct processing steps, including (1) vicariously sharing others' internal states, (2) explicitly considering (and perhaps understanding) others' states and their sources, and (3) expressing motivation to improve others' experiences (e.g. by reducing their suffering). Together, these components of empathy support our abilities to cooperate on everything from hunting trips to the development of scientific theory (Tomasello, 2009), and motivate us to protect each others' well being (Batson, 2011; de Waal, 2008).

Given empathy's enormous importance, psychologists of all stripes have developed new tools and techniques, ranging from time data in infants (Hamlin, Wynn, & Bloom, 2007; Thomsen, Frankenhuys, Ingold-Smith, & Carey, 2011) to cross-species comparative studies (Flombaum & Santos, 2005; Silk, Brosnan, Vonk, Henrich, Povinelli, Richardson, et al., 2005), to explore empathic abilities. Human neuroscience has been no different. Since the spread of tools like fMRI for measuring task-related brain activity two decades ago, researchers have called for these tools to be used in characterizing the neural bases of empathy, and many more researchers have answered this call. The resulting avalanche of data has clarified some of the myriad ways in which perceivers (individuals paying attention to, thinking about, or responding to another person) represent the experiences of social targets (individuals who are the focus of perceivers' attention).

Here, we don't aim to exhaustively review this vast literature, but rather to offer a three-part survey and glimpse of the future. The first section will describe extant neuroscience work on empathy, which has largely focused on localizing and characterizing the neural systems underlying two

components of empathy: experience sharing (perceivers' tendency to vicariously experience targets' sensorimotor, visceral, and affective states) and mental state attribution (perceivers' explicit consideration of targets' internal states), each of which has been explored by dozens of studies. This work has provided a powerful, mechanistic snapshot of some features of empathy.

The second section will explain why extant neuroscientific models of empathy remain incomplete, and as such, this domain of research is at a critical turning point. This is because a description of individual social cognitive processes—the “pieces” that make up empathy—is far removed from a holistic picture of how the human brain puts those pieces together, and allows perceivers to understand and respond to targets. This is not always appreciated in social cognitive neuroscience: researchers often treat processes such as experience sharing and mentalizing as though they were separate “processing streams” operating in isolation.

The third and last section of this chapter will focus on a critical shift in the field away from this modular view of empathy, and suggest some direction for future research. Here, we will describe how the first stage of empathy research is now giving way to a second stage that focuses on the **interactions** between multiple cognitive and neural mechanisms that constitute empathy, especially when perceivers encounter complex, ecologically valid social cues. This second stage of evolution is ongoing, and the way it plays out will determine the course of research on the neuroscience of empathy—and the issues this field will be able to address—in the coming decades.

A tale of two systems

Understanding and responding to others' internal states are enormously complex tasks. Luckily, perceivers have access to a number of methods for accomplishing them. They can stereotype social targets (Devine, 1989; Quadflieg, Turk, Waiter, Mitchell, Jenkins, & Macrae, 2009), project their own internal states onto targets (Gilovich, Medvec, & Savitsky, 2000; Ross, Greene, & House, 1977), apply analysis of variance to others' behaviors to derive underlying traits and preferences (Kelley, 1973), and avail themselves of any number of other social cognitive “tools” (Ames, 2004). That said, the lion's share of neuroscientific research on empathy has focused on two of these tools—experience sharing and mental state attribution. Neuroscientists have explored these processes and their underlying neural systems through starkly independent lines of research. Here, we will discuss each of these empirical programs in turn.

Experience sharing

The first line of research deals with the mechanisms through which one person comes to vicariously experience others' internal states. Psychologists and neuroscientists posit that experience sharing occurs because perception (e.g. of a target in pain) and experience (e.g. a perceiver feeling pain themselves) are deeply linked, and as such observing targets will naturally cause perceivers to take on those targets' states (Dijksterhuis & Bargh, 2001; Preston & de Waal, 2002). Perception-experience coupling is a centuries-old idea in philosophy (Smith, 1790/2002), and more recently has been supported by observations that perceivers indeed take on the postures (Chartrand & Bargh, 1999), facial expressions (Dimberg, Thunberg, & Elmehed, 2000), autonomic arousal (Vaughan & Lanzetta, 1980), and moods (Neumann & Strack, 2000) that they observe in others. In many ways, the idea of experience sharing follows from the more general theory of embodied cognition, which posits that concepts related to physical states (including, presumably, those of other people) are processed through sensory and motor representations (Barsalou, 2008; Decety, 1996; Kosslyn, Thompson, & Alpert, 1997; Niedenthal, Barsalou, Ric, & Krauth-Gruber, 2005; Zaki, Davis, & Ochsner, 2012).

Over the last 20 years, neuroscientists have characterized several regions of the human brain that exhibit a property consistent with experience sharing, which we will refer to as neural resonance. These regions respond to both perceivers' experience of a state and to their observation of targets experiencing that same state. As it turns out, neural resonance is widespread, and its localization depends on the type of internal state perceivers experience or observe. For example, perceivers engage the putative "mirror neuron system," encompassing premotor, inferior frontal, and inferior parietal cortex (Rizzolatti & Craighero, 2004), both when executing and observing motor acts. When experiencing and observing non-painful touch, perceivers engage somatosensory cortex (Keysers, Kaas, & Gazzola, 2010; Keysers, Wicker, Gazzola, Anton, Fogassi, & Gallese, 2004). When experiencing pain and observing targets in pain, perceivers also engage somatosensory cortex (Avenanti, Buetti, Galati, & Aglioti, 2005), but additionally recruit activity in regions related to the interoceptive and affective components of pain, including the anterior insula and anterior cingulate cortex (Jackson, Meltzoff, & Decety, 2005; Morrison, Lloyd, di Pellegrino, & Roberts, 2004; Ochsner, Zaki, Hanelin, Ludlow, Knierim, Ramachandran, et al., 2008; Singer, Seymour, O'Doherty, Kaube, Dolan, & Frith, 2004). Newer data suggest that even the hippocampus and posterior medial frontal cortex exhibit resonant properties during action imitation (Mukamel, Ekstrom, Kaplan, Iacoboni, & Fried, 2010). Hereafter, we will refer to brain regions that exhibit neural resonance as the experience sharing system (ESS), with the understanding that this is a loose, functional definition, and not one based on cytoarchitectonic properties or connectivity.

Regardless of the specific states being observed and experienced, neural resonance has generated a great deal of excitement, for at least two reasons. First, resonance has been put forward as the likely neural basis of shared representations. Second, resonance often has been nominated as the primary driver of empathy (Gallese & Goldman, 1998; Gallese, Keysers, & Rizzolatti, 2004).

The first of these claims is plausible and well supported. Neural resonance is highly consistent (Keysers & Gazzola, 2009; Rizzolatti & Sinigaglia, 2010) across studies and can be modulated by the same factors that modulate experience sharing, such as social context and perceiver motivations (Singer, Seymour, O'Doherty, Stephan, Dolan, & Frith, 2006; Xu, Zuo, Wang, & Han, 2009). Furthermore, one criticism often leveled at work on the ESS is that voxels represent relatively large patches of neural "real estate," and as such it is difficult to know whether neural resonance findings actually reflect overlapping activation in cellular populations or the activation of distinct populations that co-exist within single voxels. One extant study has addressed this concern by using multivariate techniques that hone in on multi-voxel patterns of activation while perceivers experienced pain and observed targets in pain. This relatively sensitive measure replicated the main finding of neural resonance across the two conditions (Corradi-Dell'Acqua, Hofstetter, & Vuilleumier, 2011).

The second of these claims—that neural resonance is the primary mediator of empathy—is much less well supported. This is because virtually all studies of neural resonance focus on observation and experience of relatively "low-level" states that include strong sensorimotor and visceral components, such as pain, disgust, motor intentions, and facial expressions. However, empathy involves sharing not only such low-level states with targets, but also sharing "higher level" affective states and understanding the sources of those states. Critically, high level states are often irreducible to lower level visceral or sensorimotor states; for example, the identical motor program of pushing someone could be employed for the very different high level purposes of starting a fight or saving someone from an oncoming bus (Jacob & Jeannerod, 2005). Furthermore, there are many instances in which a target's state diverges from that of a perceiver (e.g. when a target falsely believes something that a perceiver does not or is trying to hide or control expression of their true

feelings); in these cases, assuming one's own internal states are shared by a target can hinder interpersonal understanding (Epley, Keysar, Van Boven, & Gilovich, 2004; Gilovich et al., 2000).

Mental state attribution

Errors arising from imputing one's own internal states onto others, in fact, spurred early research in a very different tradition: the study of so-called "theory of mind." Since Premack & Woodruff's (1978) pioneering work with chimpanzees, scientists have studied the ability of humans (and some other animals) to ascribe unique mental states to others, and to utilize inferences about mental states during social interactions (an ability we will refer to as mental state attribution). Mental state attribution, in various forms, has been a major topic of research for decades, with special attention being paid to its developmental trajectory (Flavell, 1999), and its breakdown in autism spectrum disorders (Baron-Cohen, Leslie, & Frith, 1985).

Cognitive neuroscience research on mental state attribution over the last 15 years has borrowed a number of paradigms from these developmental and clinical traditions, usually asking perceivers to draw inferences about the beliefs, knowledge, intentions, and emotions of others based on written vignettes, pictures, or cartoons. Related work has adapted social psychological paradigms on person perception by asking perceivers to judge the stable traits (as opposed to transient states) of themselves and of targets. Regardless of the type of judgment being made about others or the medium in which target cues are presented, such tasks produce a strikingly consistent pattern of activation in a network that includes medial prefrontal cortex (mPFC), temporoparietal junction (TPJ), posterior cingulate cortex (PCC), and temporal poles. As with the ESS, we will refer to this set of regions as the mental state attribution system (MSAS), understanding that this categorization is loose and functional (for more descriptions of the MSAS and its functions, see Baron-Cohen, Ring, Wheelwright, Bullmore, Brammer, & Simmons, 1999; Castelli, Frith, Happé, & Frith, 2002; Fletcher, Happe, Frith, Baker, Dolan, Frackowiak, et al., 1995; Goel, Grafman, Sadato, & Hallett, 1995; Mitchell, 2009a; Mitchell, Heatherton, & Macrae, 2002; Ochsner, Knierim, Ludlow, Hanelin, Ramachandran, Glover, et al., 2004; Olsson & Ochsner, 2008; Peelen, Atkinson, & Vuilleumier, 2010; Saxe & Kanwisher, 2003). The specific roles of these cortical regions are, it seems, not limited to MSA-related computations. For example, the TPJ is likely related to orienting attention based on exogenous cues (Corbetta, Patel, & Shulman, 2008; Mitchell, 2008), the PCC's position as a convergence point for both sensory and motor information may support a role in assessing the salience of social stimuli (Vogt, Vogt, & Laureys, 2006), and the mPFC is often engaged by making non-social decisions under conditions of uncertainty (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006). Overall, the MSAS likely supports a suite of sub-processes that underlie a broader ability to "project" one's self into distal scenarios or points of view (including the past, future, and uncertain or counterfactual concepts, as well as targets' non-observable mental states; see Buckner, Andrews-Hanna, & Schacter, 2008; Mitchell, 2009b; Spreng, Mar, & Kim, 2009).

Isolated systems as a red herring

At first blush, it may seem that experience sharing and mental state attribution should be functional cousins, intimately linked as processes that support the broader construct of empathy. A close look reveals that—at least at the level of the brain—there is a striking lack of family resemblance, however. As readers may have noticed, the brain regions making up the ESS and the MSAS are almost completely non-overlapping (Figure 13.1). This dissociation holds up under meta-analytic scrutiny: studies engaging one system rarely concurrently engage the other (van Overwalle & Baetens, 2009). Even within individual studies, the types of cues typically engaging one system often do not

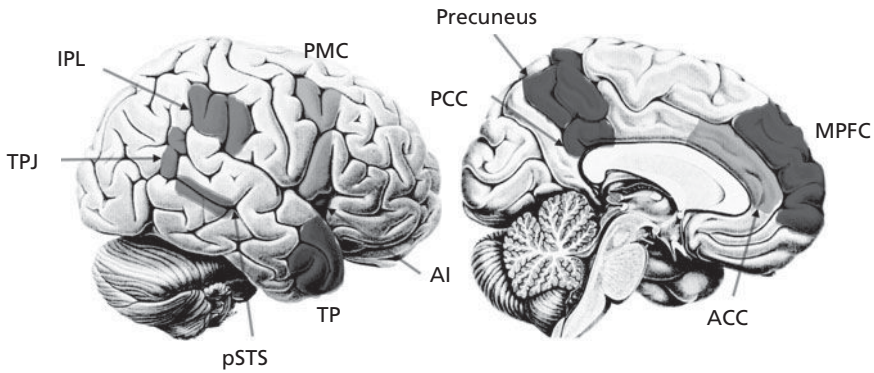


Figure 13.1 Brain regions comprising the ESS (light gray) and MSAS (dark gray). IPL = inferior parietal lobule; TPJ = temporoparietal junction; pSTS = posterior superior temporal sulcus; TP = temporal pole; AI = anterior insula; PMC = premotor cortex; PCC = posterior cingulate cortex; ACC = anterior cingulate cortex; mPFC = medial prefrontal cortex.

engage the other. Specifically, the ESS is often responsive to sensorimotor cues (e.g. facial expressions of emotion) that suggest internal states, whereas the MSAS is more responsive to *contextual* cues that describe the likely sources of those states (Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007; Wheatley, Milleville, & Martin, 2007; Zaki, in press).

Furthermore, the ESS and MSAS sometimes “compete” for control over behavior. For example, perceivers asked **not** to imitate the movements of targets attenuate activity in the mirror neuron system, but increase activity in the mPFC and TPJ (Brass, Ruby, & Spengler, 2009). Similarly, when sensorimotor and contextual cues about a target’s emotion conflict (imagine, for instance, seeing someone crying, but understanding that he has just won an Olympic gold medal), perceivers’ can rely on either type of cue in judging what they believe a target truly feels (Aviezer, Hassin, Ryan, Grady, Susskind, Anderson, et al., 2008; Carroll & Russell, 1996). Recently, we found that, in such situations, reliance on sensorimotor cues engaged activity in the ESS and dampened activity in the MSAS, whereas reliance on contextual cues produced the opposite pattern of activity (Zaki, Hennigan, Weber, & Ochsner, 2010).

Thus, the ability to empathically connect with and understand another person seems to fractionate into two disparate social “information processing streams,” supported by two dissociable neural systems. Yet ostensibly, both of these processes and neural systems serve the same ends: understanding and sharing targets’ internal states. If this is true, then what specific role does each system play in supporting empathy more broadly? Typically, reviews of this literature hold that these processes provide relatively independent contributions to empathy (Decety & Jackson, 2004; Shamay-Tsoory, 2010; Singer, 2006). Others suggest avenues through which these processes could interact with each other (Keysers & Gazzola, 2007; Uddin, Iacoboni, Lange, & Keenan, 2007).

However, until recently it was difficult to answer questions about whether and how multiple neural systems combine to support empathy, because empathy—as a multi-component phenomena—was rarely studied in neuroscience. Instead, as the review of extant data above suggests, the “first stage” of empathy research focused on characterizing single processes by engaging them in relative phenomenological isolation through the use of highly controlled unimodal, static, and artificial social cues and tasks (e.g. viewing pictures of actors posing a canonical emotional expression, imitating isolated target movements, or answering questions about vignettes describing

mental states). The reasons for beginning with such paradigms were powerful; so little was known about the neural bases of empathy that maximal experimental control was critical to gaining any insights about its constituent processes. Indeed, the control offered by simplified social paradigms was foundational in allowing scientists to build a functional architecture of empathy's building blocks.

This approach, however, also had the side effect of stacking the deck in favor of a viewing empathy's sub-processes as isolated. This is because the tasks and stimuli used to study the ESS and MSAS were typically designed to engage these processes in isolation. For example, when viewing a context-free picture of an emotional facial expression, perceivers have access only to sensorimotor cues, and as such are likely to draw heavily on experience sharing, especially because studies of the ESS rarely require perceivers to draw explicit inferences about target states. On the other hand, studies of mentalizing ask perceivers to draw just such inferences, based on heavily contextualized stimuli (e.g. vignettes describing the sources of targets' false beliefs). In other words, studies of the ESS and MSAS often employ tasks and stimuli that are "optimized" to each system and its relevant cognitive process. As such, it is unsurprising that perceivers respond by deploying the system called on by the experimental setting, in a manner consistent with modular separation between experience sharing and mental state attribution.

These differences between tasks suggest that the historical division between studies of the ESS and MSAS is both helpful and unhelpful to understanding empathy. On the one hand, it is useful to the extent that a careful approach to exploring the specific contexts in which each system is engaged can provide useful model of when and how social cues will trigger different forms of information processing. On the other hand, this approach is unhelpful if focusing on the ESS or MSAS in isolation leads to overly constrained theories of empathy (Zaki & Ochsner, 2012). Even worse, models that draw a bright line between experience sharing and mental state attribution may be overlooking a potentially central stimulus and task confound. As we will discuss below (see "Coactivation of ESS and MSAS"), the neural systems underlying these processes may be responsive to specific perceiver goals: for example, the ESS is engaged when perceivers attend to how a target is expressing an emotion, whereas the MSAS is engaged by perceivers' attention to *why* targets feel the emotions they are expressing. Furthermore, differing classes of social stimuli likely draw perceiver attention naturally towards the "how" or "why" of targets' actions and expressions, and these are typically the types of cues that studies of the ESS and MSAS have employed. Similarly, different tasks, e.g. imitating a targets' facial expression vs. judging how a target likely feels based on that expression—orient perceivers towards different goals and engage different neural systems, even when based on nearly identical stimuli (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003; Mitchell et al., 2002).

"First stage" empathy research typically—though not always—divided programs of research along more than one of these dimensions. Studies of the ESS typically used low level sensorimotor cues and either passive viewing or imitation tasks, whereas studies of the MSAS often used more contextualized cues and drew perceivers' attention towards explicit judgments of target states. Do the resulting findings necessarily mean that experience sharing and mental state attribution are isolated in natural contexts? It is quite difficult to answer this question, because we do not know whether differences in neural activity in studies of the ESS and MSAS reflect stimulus type, task, or perceiver attentional set, as opposed to true distinctions between information processing streams.

Critical here is the fact that the social cues perceivers encounter outside the laboratory are often substantially different than those employed by the lion's share of extant research. Specifically, "real-world" social cues are typically dynamic (unfolding over time), multimodal (including concurrent sensorimotor and contextual information), and contextually embedded (such that

interpretations of any one cue are often altered or constrained by other cues or a perceiver's prior knowledge; see Keysers & Gazzola, 2007; Zaki & Ochsner, 2009). The gulf separating laboratory and naturalistic social cues would not be problematic if these cues produced the same patterns of brain activity, and only differed in, for example, the intensity of this activity (i.e. differing quantitatively). However, early evidence suggests this is not the case. Indeed, naturalistic social information seems to produce patterns of information processing and brain activity that differ qualitatively from those produced by simplified cues like those used in typical social cognitive neuroscience studies, including engaging both the ESS and MSAS, and producing interactions between these systems.

This is important, in part, because it is under-acknowledged in much of the neuroscience literature on empathy. Resulting theoretical models, in turn, may over-emphasize the ability to understand empathy as a whole based on tasks examining isolate pieces of typical social experiences. By way of analogy, this may be something like drawing inferences about the way that the brain processes the sound of an orchestra based on data describing how the brain processes the sound of each individual instrument, ignoring the unique types of information (e.g. harmonies across instruments) that emerge at the orchestral level in the real world stimulus of interest (Zaki, under revision).

Theories emphasizing the dissociability of the MSAS and ESS run the risk of either missing or glossing over this complexity, and as a consequence, formulating models of empathy that rest too heavily on single processes. For example, two competing and well-known theories have claimed that interpersonal cognition can be largely localized to either the ESS or MSAS (Gallese et al., 2004; Saxe, 2005). The resulting debate, while provocative, is probably misguided, because each side bases its argument on evidence derived from studies examining only one piece of the larger social puzzle.

Putting the pieces together

So far, we have chronicled the work researchers have done in characterizing the neural bases of two empathic sub-processes—experience sharing and mental state attribution—and described some conceptual limitations that hinder the ability of descriptions of single social cognitive “pieces” to translate into descriptions of empathy as it likely operates in more ecologically valid contexts (cf. Neisser, 1980).

This second point is not meant to discredit work on single social cognitive processes. Quite the opposite—such research is not only important, but also constitutes the only reasonable starting point for building a neuroscience of empathy. That said, we (Zaki, in press; Zaki & Ochsner, 2009) and others (e.g. Keysers & Gazzola, 2007) have advocated for following this research with a “second stage” of work focusing not on single processes in isolation, but on how perceivers put these pieces together, by deploying multiple, interactive empathic processes when encountering complex social cues.

Luckily, this second stage is well underway. Largely in the last 3 years, researchers have updated their approach to examine the brain's response to just the type of complex social information we have described above. This work capitalizes on first stage characterizations of the ESS and MSAS to study how these systems respond when pieces of social information (e.g. dynamic biological movement and linguistic cues about beliefs or emotions) are joined to form a coherent whole.

This work has produced a sea change in the way neuroscientists view empathy. Instead of conceiving of experience sharing and mental state attribution as isolated social cognitive processing streams, we now have a picture of these processes as intimately tied in at least 3 ways: As reviewed below (1) the ESS and MSAS are concurrently engaged by naturalistic social cues, (2) these neural

systems become functionally coupled with each other during complex social cognitive tasks, and (3) activation of both of these systems predict empathy-related outcomes, including accuracy about targets' internal states and perceivers' motivation to engage in prosocial behavior towards targets.

Coactivation of the ESS and MSAS

Early data led to the suggestion that the ESS and MSAS were fundamentally dissociable, but—as mentioned above—this was based on stimuli and paradigms designed to isolate single social cognitive processes. Outside the laboratory, social targets more often than not present us with a barrage of multimodal social cues that unfold over time (e.g. a friend looks uncomfortable, then reveals that she has just lost her job, and then leans forward and begins crying). Such cues tap all of our social capacities simultaneously and demand that we integrate over many social signals in forming a coherent representation of targets' internal states. Intuitively, we might expect that such demands would engage multiple social cognitive processes and neural systems.

Consistent with this, several studies combining complex, dynamic social stimuli with the requirement for explicit inferences about targets' states (requirements often present in typical social interactions) have consistently demonstrated concurrent engagement of both areas within the ESS and MSAS. In many cases, these studies also help to reveal the specific contextual triggers that produce such coactivation. For example, watching videos of targets executing motor acts engages areas within the ESS involved in sharing motor intentions; if these videos are further paired with demands to draw explicit inferences about targets' intentions—or situational cues drawing attention to targets' likely intentions—they also engage areas in the MSAS (de Lange, Spronk, Willems, Toni, & Bekkering, 2008; Spunt, Satpute, & Lieberman, 2010; Wheatley et al., 2007). Similarly, engaging in a joint attention task with a target engages regions within both of these neural systems (Redcay, Dodell-Feder, Pearrow, Mavros, Kleiner, Gabrieli, et al., 2010). Together, these data suggest that areas within the ESS may engaged relatively automatically by dynamic social stimuli (e.g. moving social targets), but that requirements to further digest the internal states implied by targets' movements brings the MSAS online as well (Spunt & Lieberman, 2011).

These patterns of coactivation translate to emotion perception as well. For example, when perceivers view videos of targets expressing emotions, they typically engage both areas within the ESS and MSAS (Wolf, Dziobek, & Heekeren, 2010; Zaki, Weber, Bolger, & Ochsner, 2009). Furthermore, the system that comes online most strongly under such circumstances may depend on perceivers' inferential goals. A recent study elegantly demonstrated this point: when attending to the way targets express their emotions (e.g. through laughing), perceivers prominently engaged the ESS—and especially regions involved in sharing motor intentions—whereas attending to the sources of target emotions (e.g. hearing a good joke), perceivers most strongly engaged the MSAS (Spunt & Lieberman, 2013). Such findings not only provide us with a more holistic picture of coactivation in these systems, but also refine our understanding of the specific social sensitivities exhibited by these neural systems.

These data make an important point about how theories of empathy should discuss prior data. That is, the fact that the ESS and MSAS can be dissociated using simplified stimuli and tasks does not necessitate, or even imply, that those systems are dissociable in the majority of social contexts. In fact, studies employing naturalistic methods suggest that the demands of most social situations would engage these systems—and the processes they underlie—simultaneously. This probability motivates a shift away from an “either / or” argument about whether the MSAS or ESS is central to empathy, and towards a “when and how” approach to better discriminating the situations likely to engage one or both systems.

Functional coupling between systems

In addition to being engaged together, a parsimonious account of empathy might posit that processes such as experience sharing and mental state attribution should intricately interact during naturalistic social cognition. For example, understanding the sources behind a target's likely internal states (e.g. that he has just won a gold medal) could cause perceivers to vastly reinterpret that target's sensorimotor cues (e.g. crying). Presumably, this efficient use of multiple pieces of social information could be instantiated through communication between the ESS and MSAS.

Consistent with this approach, a number of studies have documented functional coupling between the ESS and MSAS during social cognitive tasks. For example, classic work on neural resonance demonstrates that areas within the "pain matrix" (especially the anterior insula and anterior cingulate cortex) are engaged both when targets experience pain themselves and when they observe targets in pain (Singer et al., 2004). However, this does not mean that these regions are performing identical computations during both pain perception and experience. On the contrary, the interpretation of a single region's activity depends on the other regions with which that region communicates during a given task. Our own group has examined this idea within the context of empathy for pain. We found that the ACC and AI were engaged during both experience and observation of pain (Ochsner et al., 2008), but that these regions demonstrated very different patterns of connectivity across these tasks: during observation, but not experience, ACC and AI became functionally coupled with areas within the MSAS (Zaki, Ochsner, Hanelin, Wager, & Mackey, 2007). Similar connectivity patterns also apply to the experience and observation of disgust (Jabbi, Bastiaansen, & Keysers, 2008). Together, these data suggest that, during experience sharing tasks, neural resonance—shared activity for self and other experience—may depend on communication with regions involved in mental state attribution.

Other studies have tested the other side of this equation: examining the connectivity of areas in the MSAS during an explicit social inference task. For example, Lombardo, Chakrabarti, Bullmore, Wheelwright, Sadek, Suckling, et al. (2010) asked perceivers to draw inferences about their own preferences and those of targets. Both of these conditions engaged many regions classically making up the MSAS, including the mPFC, PCC, and TPJ. Interestingly, during both types of inference, the mPFC and TPJ were also functionally connected with many regions in the ESS, regardless of whether participants answered questions about themselves or social targets. This suggests that even relatively simple inferences about targets may require communication between regions involved in drawing such inferences and regions involved in sharing of intention and affect with targets.

Connectivity can also be studied *across* perceivers and targets. For example, Schippers, Roebroek, Renken, Nanetti, & Keysers (2010) asked targets to manually pantomime simple actions (*à la charades*) while being scanned using fMRI; perceivers were later scanned while guessing what gesturers were attempting to communicate. Using an innovative analysis, the researchers demonstrated that activity in targets' motor cortex while they executed a gesture predicted activity in perceivers' motor cortex while they perceived those gestures. Interestingly, however, targets' motor activation also predicted activity within perceivers' MSAS—specifically the mPFC—suggesting that perceivers process targets' intentions using both the ESS and MSAS. Further, communication between the mPFC and areas within the ESS are modulated by perceivers' intentions to actively guess what targets are pantomiming vs. passively viewing target actions (Schippers & Keysers, 2011), again suggesting that situational and motivational context critically affect the interplay between empathic sub-processes.

Predicting social cognitive outcomes

When mapping the neural architecture of any complex cognitive process, a key concern is that brain activity in a given region does not actually index the computational process a researcher is interested in. Empathy is not excepted from this issue, and based in imaging data alone, it is difficult to know exactly what engagement of, for example, ESS regions during a shared experience task actually means psychologically.

This has been especially problematic because of the distance that has historically separated psychological and neuroscientific approaches to empathy. Social psychological approaches—perhaps not surprisingly—lean heavily on behavior to indicate the operation of empathic processes. For example, perceivers' accuracy for targets' internal states can serve as an indicator of how much perceivers engage both mentalizing and experience sharing (Ickes, 1997; Levenson & Ruef, 1992; Tetlock & Kim, 1987), whereas stereotyping or derogation of targets can index the absence of these processes (Devine, 1989; Harris & Fiske, 2007). Similarly, perceivers' choices to engage in prosocial behavior can serve as an index of their concern for targets' well being (Batson, 2011).

By contrast, neuroimaging studies of empathy—especially during the “first stage”—concentrated far less on behavioral outcomes, and more on relationships between stimuli and brain activity. For example, perceivers might be scanned while observing targets in pain or making guesses about targets' intentions, and related brain activity would be interpreted as relevant to the empathic sub-process that task putatively engages. In almost all cases, these paradigms do not produce variance in behavior, either because they required no responses from perceivers (as in many passive experience sharing tasks) or employed very simple social inference tasks that produce near perfect accuracy.

This precluded neuroimaging studies of empathy from mapping brain activity directly on to behavior, reducing the ability of researchers to draw maximally strong inferences about neuroimaging results. For example, although the ESS is engaged during observation of pain, in the absence of brain-behavior relationships, it is difficult to know whether this activation actually tracks with experience sharing, or instead tracks concurrent, but less interesting, processing step (e.g. remembering one's own painful experiences, desire to escape the discomfort of observing suffering, a perceiver's attempt to distract himself from viewing the unpleasant stimulus). Individual difference correlations (Jabbi, Swart, & Keysers, 2007) and lesion studies (Shamay-Tsoory, Aharon-Peretz, & Perry, 2009) provide a partial remedy to this concern, but cannot replace the utility of brain-behavior links.

Other domains within cognitive neuroscience have fruitfully studied brain-behavior correlations. Notably, memory researchers used the subsequent memory paradigm to link encoding-related activation in the medial temporal lobe and inferior frontal cortex to successful retrieval of memoranda (Brewer, Zhao, Desmond, Glover, & Gabrieli, 1998; Wagner, Schacter, Rotte, Koutstaal, Maril, Dale, et al., 1998). The “second stage” of empathy research has picked up this trend, by relating activity in the MSAS and ESS to subsequent social behaviors, including accuracy for social information and subsequent prosocial behavior.

With respect to accuracy, early work “piggybacked” on the original subsequent memory paradigm to examine whether and how brain activity would predict accurate recall for social—as opposed to non-social—information. A spate of studies demonstrated that MSAS activity when perceivers encounter socially relevant stimuli (e.g. trait adjectives) predicted successful retrieval of this information, but only when perceivers were drawing social inferences about those stimuli (e.g. how much an adjective described a social target, as opposed to how many vowels it contained) (Hasson, Furman, Clark, Dudai, & Davachi, 2008; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004; Mitchell, Macrae, & Banaji, 2004). A later study took this approach into a more naturalistic

context, demonstrating that reliable patterns of activity within both MSAS and ESS areas predicted the accuracy with which perceivers recall targets' descriptions of autobiographical events (Stephens, Silbert, & Hasson, 2010).

Our group has examined brain-behavior correlations in the affective domain, by studying the neural correlates of accurate inferences about targets' emotions based on naturalistic social cues (Ickes, 1997; Levenson & Ruef, 1992; Zaki, Bolger, & Ochsner, 2008; Zaki & Ochsner, 2011). In our studies, perceivers watch videos of targets describing emotional events, and continuously rate how positive or negative they believed targets feel. Importantly, targets themselves had previously rated their emotions at each moment using the same scale perceivers employed. This allows us to quantitatively operationalize interpersonal accuracy as the correlation between perceivers' ratings of targets emotions and targets' self-ratings. Using this approach, we have demonstrated that accuracy is predicted by activity in regions in both the MSAS and ESS (Harvey, Zaki, Lee, Ochsner, & Green, *in press*; Zaki et al., 2009).

Finally, a small set of studies has examined brain activity related to the use of mental state information during game theoretic decision-making. Although not measuring accuracy *per se*, these tasks offer the attractive possibility of formally modeling the use of mental states in interpersonal strategizing. For example, in both the "work / shirk" and "beauty contest" games, perceivers must strategically infer what others will think in order to maximize their own gains. In both of these games, perceivers level of social inference (e.g. how much their decisions reflect thinking about others' minds) can be quantified; and, in both cases, activity in the MSAS—and specifically the mPFC—tracks with this measure (Coricelli & Nagel, 2009; Hampton, Bossaerts, & O'Doherty, 2008). This strengthens the inference that the MSAS directly tracks with the insightful, task-related use of others' mental states during social interactions.

A second growing literature has focused on brain-behavior correlations in another domain: using brain activity to predict prosocial behaviors such as sharing resources and helping social targets. The motives behind prosocial behavior have been the topic of a high profile debate among social psychologists. Interestingly, this debate can be recast along the dimensions of experience sharing and mental state attribution: whereas Cialdini and colleagues (Cialdini, Brown, Lewis, Luce, & Neuberg 1997; Cialdini, Schaller, Houlihan, Arps, Fultz, & Beaman, 1987) suggested that prosocial behavior stemmed from a sense of "oneness" or overlapping identity with targets (akin to experience sharing), Batson and colleagues (Bateson, 1991, 2011; Bateson et al., 1991) countered that a specific form of other oriented cognition (akin to mental state attribution) was the stronger driver of prosocial behavior.

Which one of these mechanisms supports prosociality? Neuroscience can provide converging evidence through which this question can be addressed, by examining the extent to which activity in the MSAS and ESS predicts later prosocial acts. Like so many features of second stage neuroscience work on empathy, the emerging answer seems to be that both systems are involved, in a context-dependent manner. For example, ESS activity consistent with perceivers' sharing of targets' pain (Hein, Silani, Preuschoff, Batson, & Singer, 2010; Masten, Morelli, & Eisenberger, 2011) and reward (Harbaugh, Mayr, & Burghart, 2007; Zaki, Lopez, & Mitchell, 2013; Zaki & Mitchell, 2011) predicts perceivers' willingness to make costly decisions that help those targets. In other cases, activity in the MSAS (especially the mPFC) when perceivers consider targets' internal states predicts their later willingness to act prosocially (Morelli, Rameson, & Lieberman, 2012; Waytz, Zaki, & Mitchell, 2012).

The specific contextual factors that determine when activity in the MSAS or ESS will best predict prosocial behavior remain relatively unexplored. Future work should address this issue, and examine whether prosocial behavior prompted by experience sharing and mental state attribution,

respectively, differ in their subjective or behavioral features. Nonetheless, the small, but growing literature on this topic clearly provides evidence consistent with both sides of the psychological debate: under at least some conditions, it seems that both of these processes can drive prosocial motivations.

Conclusions and future directions

For all of our impressive mental firepower, most humans (including both authors of this chapter) would not last a week alone in the wild. But put us together in a group, and we can survive just about anything. Why? In simple terms, it is because we are built for other people, in more ways than one. Behaviorally, our species has thrived on coordinated interpersonal actions. Psychologically, we are equipped with myriad affective and cognitive mechanisms perfectly suited to understanding and relating to other minds. Empathy and all of its components are foundational to who we are and why we succeed as a species.

Although empathy has been a perennial topic among philosophers, the neuroscience of empathy is in its teenage years. Given its immaturity, the rapid evolution of this domain of research is especially impressive. Here we have chronicled two of these evolutionary “stages.” In the first stage, researchers characterized the neural systems supporting two major empathic sub-processes: the ESS, which is involved in sharing targets’ sensorimotor and visceral states, and the MSAS, which is involved in perceivers’ explicit inferences about targets’ states.

This work was hugely important in building a functional architecture of empathy. However, it was also hamstrung by two important problems. First, in mapping the neural bases of empathic sub-processes, researchers necessarily began by using highly simplified non-naturalistic social cues, and this sometimes led to overly constrained models of empathy as comprising a number of “pieces” that operated in relative isolation. Secondly, first stage empathy research in neuroscience rarely related brain activity to observable social behaviors, making it difficult to draw direct conclusions about the functional role of the ESS and MSAS.

The “second stage” of this program has begun to remedy these issues. Critically, however, it has not overwritten the first stage, but rather built on the important insights of earlier work. Specifically, it has capitalized on first-stage descriptions of the ESS and MSAS to further demonstrate (1) that these systems are concurrently engaged by naturalistic, multimodal social cues, (2) that they interact with each other when processing such stimuli, and (3) that their engagement can predict subsequent social-behavioral outcomes such understanding targets’ internal states and motivations to help targets. This work provides an integrative view of empathy as tapping multiple, functionally connected sets of brain regions to translate complex social cues into inferences about others’ internal states. Further, second stage research has highlighted the context-dependent nature of empathy: depending on situational features, the same social cues can engage very different patterns of activity across the ESS and MSAS, and the activity of these systems can differentially predict subsequent social behaviors (cf. Hein & Singer, 2008). Overall, the second stage of empathy research has refined and integrated models of isolated empathic sub-processes into more holistic accounts of an integrated “system” of processes that perceivers deploy flexibly based on current social goals and information.

This summary begs the question of what a “third stage” of neuroscience research on empathy might bring. Although this is difficult (if not impossible) to predict, the insights garnered by the first two stages of work suggest some exciting possibilities, two of which we will mention here. First, extant work has yet to capture—in any meaningful way—a central feature of social encounters: the fact that perceivers themselves are also usually targets, and visa-versa. Unless they are

watching television, perceivers rarely observe targets without themselves being observed. As such, much of perceivers' ongoing social cognitive labor entails iteratively sampling their effect on targets (wondering, e.g. "Does she know I'm paying attention? How is what I'm saying now affecting him?"), and adjusting their behavior accordingly (Neisser, 1980; Schilbach, 2010). Future work should examine whether these unique features of social interactions are subserved by the same networks of brain regions that are involved in observing non-interactive targets, or whether extant work may have yet to chart the neural bases of some critical features of everyday empathy.

A second exciting avenue for future work lies in the use of quantitative models to formally describe the role of neural systems in producing social behavior. First and second stage research on empathy have equipped us with reliable insights about the neural signatures of processes such as mental state attribution, and we can now use these signatures to directly model the relationship between these processes and "downstream" inferences, decisions, and behaviors. This type of advance also has the potential to increase our understanding of potential parallels between social cognition and other domains, such as perceptual decision-making (Freeman, Schiller, Rule, & Ambady, 2010; Zaki, in press) and reinforcement learning (Behrens, Hunt, Woolrich, & Rushworth, 2008; Jones, Somerville, Li, Ruberry, Libby, Glover, et al., 2011).

The neuroscience of empathy has evolved fruitfully by consistently building on prior work to refine and improve the questions and models this field produces. So long as this trajectory continues, this field will continue growing at an amazing pace, and producing fundamental insights about the nature of our critical social abilities.

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(3): 340–53.
- Avenanti, A., Bueti, D., Galati, G., & Aglioti, S. M. (2005). Transcranial magnetic stimulation highlights the sensorimotor side of empathy for pain. *Nature Neuroscience* 8(7): 955–60.
- Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., et al. (2008). Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychology Science* 19(7): 724–32.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition* 21(1): 37–46.
- Baron-Cohen, S., Ring, H. A., Wheelwright, S., Bullmore, E. T., Brammer, M. J., Simmons, A., et al. (1999). Social intelligence in the normal and autistic brain: an fMRI study. *European Journal of Neuroscience* 11(6): 1891–8.
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology* 59: 617–45.
- Batson, C. D. (1991). *The Altruism Question: Toward A Social-psychological Answer*. Hillsdale: Lawrence Erlbaum.
- Batson, C. D. (2011). *Altruism in Humans*. Oxford: Oxford University Press.
- Batson, C. D., Batson, J. G., Slingsby, J. K., Harrell, K. L., Peekna, H. M., & Todd, R. M. (1991). Empathic joy and the empathy-altruism hypothesis. *Journal of Personality and Social Psychology* 61(3): 413–26.
- Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature* 456(7219): 245–9.
- Brass, M., Ruby, P., & Spengler, S. (2009). Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society of London, B Biological Science* 364(1528): 2359–67.
- Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1998). Making memories: brain activity that predicts how well visual experience will be remembered. *Science* 281(5380): 1185–7.
- Brothers, L. (1997). *Friday's Footprint: How Society Shapes the Human Mind*. New York: Oxford University Press.

- Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Science* 1124: 1–38.
- Carr, L., Iacoboni, M., Dubeau, M. C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences USA* 100(9): 5497–502.
- Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology* 70(2): 205–18.
- Castelli, F., Frith, C., Happe, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain* 125(Pt 8): 1839–49.
- Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology* 76(6): 893–910.
- Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.
- Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy-altruism relationship: when one into one equals oneness. *Journal of Personality and Social Psychology* 73(3): 481–94.
- Cialdini, R. B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., & Beaman, A. L. (1987). Empathy-based helping: is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology* 52(4): 749–58.
- Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron* 58(3): 306–24.
- Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences, USA* 106(23): 9163–8.
- Corradi-Dell'acqua, C., Hofstetter, C., & Vuilleumier, P. (2011). Felt and seen pain evoke the same local patterns of cortical activity in insular and cingulate cortex. *Journal of Neuroscience* 31(49): 17996–8006.
- Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Y. Munakata & M. Johnson (Eds), *Processes of Change in Brain and Cognitive Development. Attention and Performance, XXI* (Vol. 249–274). Oxford: Oxford University Press.
- Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature* 441(7095): 876–9.
- de Lange, F. P., Spronk, M., Willems, R. M., Toni, I., & Bekkering, H. (2008). Complementary systems for understanding action intentions. *Current Biology* 18(6): 454–7.
- de Waal, F. B. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual Review of Psychology* 59, 279–300.
- Deacon, T. W. (1997). *The Symbolic Species: the Co-evolution of Language and the Brain*, 1st edn. New York: W. W. Norton.
- Decety, J. (1996). Do imagined and executed actions share the same neural substrate? *Brain Research in Cognitive Brain Research* 3(2): 87–93.
- Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behaviour and Cognitive Neuroscience Review* 3(2): 71–100.
- Devine, P. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* 56(1): 5–18.
- Dijksterhuis, A., & Bargh, J. (2001). The perception-behavior Expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology* 33: 1–40.
- Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychology Science* 11(1): 86–9.
- Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology* 87(3): 327–39.
- Flavell, J. (1999). Cognitive development: Children's knowledge about other minds. *Annual Review of Psychology* 50: 21–45.

- Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., et al. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition* 57(2): 109–28.
- Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. [Comparative Study Research Support, N.I.H., Extramural Research Support, Non-U. S. Gov't Research Support, U. S. Gov't, Non-P.H.S. Research Support, U. S. Gov't, P.H.S.]. *Current Biology* 15(5): 447–52.
- Freeman, J. B., Schiller, D., Rule, N. O., & Ambady, N. (2010). The neural origins of superficial and individuated judgments about ingroup and outgroup members. *Human Brain Mapping* 31(1): 150–9.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Science* 2(12): 493–501.
- Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Science* 8(9): 396–403.
- Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology* 78: 211–22.
- Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: a comparison of theory of mind tasks. *Journal of Cognitive Neuroscience* 19(11): 1803–14.
- Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *Neuroreport* 6(13): 1741–6.
- Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature* 450(7169): 557–9.
- Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Science, USA* 105(18): 6741–6.
- Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 316(5831): 1622–5.
- Harris, L. T., & Fiske, S. T. (2007). Social groups that elicit disgust are differentially processed in mPFC. *Social Cognitive & Affective Neuroscience* 2(1): 45–51.
- Harvey, P. O., Zaki, J., Lee, J., Ochsner, K., & Green, M. F. (in press). Neural substrates of empathic accuracy in people with schizophrenia. *Schizophrenia Bulletin*.
- Hasson, U., Furman, O., Clark, D., Dudai, Y., & Davachi, L. (2008). Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. *Neuron* 57(3): 452–62.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? [Review]. *Science* 298(5598): 1569–79.
- Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron* 68(1): 149–60.
- Hein, G., & Singer, T. (2008). I feel how you feel, but not always: the empathic brain and its modulation. *Current Opinions in Neurobiology* 18(2): 153–8.
- Ickes, W. (1997). *Empathic Accuracy*. New York: Guilford Press.
- Jabbi, M., Bastiaansen, J., & Keysers, C. (2008). A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PLoS One* 3(8): e2939.
- Jabbi, M., Swart, M., & Keysers, C. (2007). Empathy for positive and negative emotions in the gustatory cortex. *NeuroImage* 34(4): 1744–53.
- Jackson, P. L., Meltzoff, A. N., & Decety, J. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. *NeuroImage* 24(3): 771–9.
- Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: a critique. *Trends in Cognitive Sciences* 9(1): 21–5.
- Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., et al. (2011). Behavioral and neural properties of social reinforcement learning. *Journal of Neuroscience* 31(37): 13039–45.

- Kelley, H. (1973). The process of causal attribution. *American Psychology* 28(2): 107–28.
- Keysers, C., & Gazzola, V. (2007). Integrating simulation and theory of mind: from self to social cognition. *Trends in Cognitive Science* 11(5): 194–6.
- Keysers, C., & Gazzola, V. (2009). Expanding the mirror: vicarious activity for actions, emotions, and sensations. *Current Opinions in Neurobiology* 19(6): 666–71.
- Keysers, C., Kaas, J. H., & Gazzola, V. (2010). Somatosensation in social perception. *Nature Reviews in Neuroscience* 11(6): 417–28.
- Keysers, C., Wicker, B., Gazzola, V., Anton, J. L., Fogassi, L., & Gallese, V. (2004). A touching sight: SII/PV activation during the observation and experience of touch. *Neuron* 42(2): 335–46.
- Kosslyn, S. M., Thompson, W. L., & Alpert, N. M. (1997). Neural systems shared by visual imagery and visual perception: a positron emission tomography study. *NeuroImage* 6(4): 320–34.
- Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition* 50(1–3): 211–38.
- Levenson, R. W., & Ruef, A. M. (1992). Empathy: a physiological substrate. *Journal of Personality & Social Psychology* 63(2): 234–46.
- Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., et al. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience* 22(7): 1623–35.
- Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex* 14(6): 647–54.
- Masten, C. L., Morelli, S. A., & Eisenberger, N. I. (2011). An fMRI investigation of empathy for ‘social pain’ and subsequent prosocial behavior. *NeuroImage* 55(1): 381–8.
- Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex* 18(2): 262–71.
- Mitchell, J. P. (2009a). Inferences about mental states. *Philosophical Transactions of the Royal Society of London, B Biological Science* 364(1521): 1309–16.
- Mitchell, J. P. (2009b). Social psychology as a natural kind. *Trends in Cognitive Science* 13(6): 246–51.
- Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences, USA* 99(23): 15238–43.
- Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2004). Encoding-specific effects of social cognition on the neural correlates of subsequent memory. *Journal of Neuroscience* 24(21): 4912–17.
- Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: the Vygotskian intelligence hypothesis. [Review]. *Philosophical Transactions of the Royal Society of London, B Biological Sciences* 362(1480): 639–48.
- Morelli, S. A., Rameson, L. T., & Lieberman, M. D. (2012). The neural components of empathy: Predicting daily prosocial behavior. *Social Cognitive and Affective Neuroscience*.
- Morrison, I., Lloyd, D., di Pellegrino, G., & Roberts, N. (2004). Vicarious responses to pain in anterior cingulate cortex: is empathy a multisensory issue? *Cognitive and Affective Behavioural Neuroscience* 4(2): 270–8.
- Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology*.
- Neisser, U. (1980). On “social knowing”. *Personality and Social Psychology Bulletin* 6(4): 601–5.
- Neumann, R., & Strack, F. (2000). “Mood contagion”: the automatic transfer of mood between persons. *Journal of Personality and Social Psychology* 79(2): 211–23.
- Niedenthal, P., Barsalou, L. W., Ric, F., & Krauth-Gruber, S. (2005). Embodiment in the acquisition and use of emotion knowledge. In L. Feldman Barrett, P. Niedenthal & P. Winkielman (Eds), *Emotion and Consciousness* (pp. 21–50). New York: Guilford Press.
- Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., et al. (2004). Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience*, 16(10): 1746–72.

- Ochsner, K. N., Zaki, J., Hanelin, J., Ludlow, D. H., Knierim, K., Ramachandran, T., et al. (2008). Your pain or mine? Common and distinct neural systems supporting the perception of pain in self and others. *Social Cognitive Affective Neuroscience* 3(2): 144–60.
- Olsson, A., & Ochsner, K. N. (2008). The role of social cognition in emotion. *Trends in Cognitive Science* 12(2): 65–71.
- Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience* 30(30): 10127–34.
- Pinker, S. (1994). *The Language Instinct*, 1st edn. New York: W. Morrow and Co.
- Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioural Brain Science* 13: 707–84.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioural Brain Science* 1: 515–26.
- Preston, S. D., & de Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioural Brain Science* 25(1): 1–20; discussion 20–71.
- Quadflieg, S., Turk, D. J., Waiter, G. D., Mitchell, J. P., Jenkins, A. C., & Macrae, C. N. (2009). Exploring the neural correlates of social stereotyping. *Journal of Cognitive Neuroscience* 21(8): 1560–70.
- Redcay, E., Dodell-Feder, D., Pearrow, M. J., Mavros, P. L., Kleiner, M., Gabrieli, J. D., et al. (2010). Live face-to-face interaction during fMRI: A new tool for social cognitive neuroscience. *NeuroImage* 50(4): 1639–47.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience* 27, 169–92.
- Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature Review of Neuroscience* 11(4): 264–74.
- 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(3): 279–301.
- Saxe, R. (2005). Against simulation: the argument from error. *Trends in Cognitive Science* 9(4): 174–9.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind”. *NeuroImage* 19(4): 1835–42.
- Schilbach, L. (2010). A second-person approach to other minds. *Nature Review of Neuroscience* 11(6): 449.
- Schippers, M. B., & Keysers, C. (2011). Mapping the flow of information within the putative mirror neuron system during gesture observation. *NeuroImage* 57(1): 37–44.
- Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flow from one brain to another during gestural communication. *Proceedings of the National Academy of Science, USA* 107(20): 9388–93.
- Shamay-Tsoory, S. G. (2010). The neural bases for empathy. *The Neuroscientist*.
- Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus vs. ventromedial prefrontal lesions. *Brain* 132(Pt 3): 617–27.
- Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., et al. (2005). Chimpanzees are indifferent to the welfare of unrelated group members. [Research Support, N.I.H., Extramural Research Support, Non-U. S. Gov’t Research Support, U. S. Gov’t, P.H.S.]. *Nature* 437(7063): 1357–9.
- Singer, T. (2006). The neuronal basis and ontogeny of empathy and mind reading: review of literature and implications for future research. *Neuroscience and Biobehaviour Review* 30(6): 855–63.
- Singer, T., Seymour, B., O’Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective, but not sensory components of pain. *Science* 303(5661): 1157–62.
- Singer, T., Seymour, B., O’Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439(7075): 466–9.

- Smith, A. (1790/2002). *The Theory of Moral Sentiments*. Cambridge: Cambridge University Press.
- Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of Cognitive Neuroscience* 21(3): 489–510.
- Spunt, R. P., & Lieberman, M. D. (2011). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. *NeuroImage* 59(1): 3050–9.
- Spunt, R. P., & Lieberman, M. D. (2013). The busy social brain: an fMRI study of cognitive load during action observation. *Psychological Science* 24(1): 80–6.
- Spunt, R. P., Satpute, A. B., & Lieberman, M. D. (2010). Identifying the What, Why, and How of an Observed Action: An fMRI Study of Mentalizing and Mechanizing during Action Observation. *Journal of Cognitive Neuroscience* 23(1): 63.
- Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Science, USA* 107(32): 14425–30.
- Suddendorf, T., & Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. [Review]. *Genetic, Social, and General Psychology Monographs* 123(2): 133–67.
- Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? [Research Support, N.I.H., Extramural Research Support, Non-U. S. Gov't Review]. *Behavioral and Brain Sciences* 30(3), 299–313; discussion 313–351.
- Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology* 52(4): 700–9.
- Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: preverbal infants mentally represent social dominance. [Research Support, N.I.H., Extramural Research Support, Non-U. S. Gov't]. *Science* 331(6016): 477–80.
- Tomasello, M. (2000). *The Cultural Origin of Human Cognition*. Cambridge: Harvard University Press.
- Tomasello, M. (2009). *Why We Cooperate*. Cambridge: MIT Press.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behaviour and Brain Sciences* 28(5): 675–91; discussion 691–735.
- Tulving, E. (2002). Episodic memory: from mind to brain. [Case Reports]. *Annual Review of Psychology* 53, 1–25.
- Uddin, L. Q., Iacoboni, M., Lange, C., & Keenan, J. P. (2007). The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends in Cognitive Science* 11(4): 153–7.
- van Overwalle, F., & Baetens, K. (2009). Understanding Others' Actions and Goals by Mirror and Mentalizing Systems: A Meta-analysis. *NeuroImage* 48(3): 564–84.
- 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(6): 909–23.
- Vogt, B. A., Vogt, L., & Laureys, S. (2006). Cytology and functionally correlated circuits of human posterior cingulate areas. *NeuroImage* 29(2): 452–66.
- Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., et al. (1998). Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science* 281(5380): 1188–91.
- Waytz, A., Zaki, J., & Mitchell, J. (2012). Response of dorsomedial prefrontal cortex predicts altruistic behavior. *Journal of Neuroscience* 32: 7646–50.
- Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding animate agents: distinct roles for the social network and mirror system. *Psychology Science* 18(6): 469–74.
- Wolf, I., Dziobek, I., & Heekeren, H. R. (2010). Neural correlates of social cognition in naturalistic settings: a model-free analysis approach. *NeuroImage* 49(1): 894–904.
- Xu, X., Zuo, X., Wang, X., & Han, S. (2009). Do you feel my pain? Racial group membership modulates empathic neural responses. *Journal of Neuroscience* 29(26): 8525–9.

- Zaki, J. (in press). Cue integration: A common framework for physical perception and social cognition. *Perspectives in Psychological Sciences*.
- Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science* 19(4): 399–404.
- Zaki, J., Davis, J. I., & Ochsner, K. (2012). Overlapping activity in anterior insula during interoception and emotional experience. *NeuroImage* 62(1): 493–9.
- Zaki, J., Hennigan, K., Weber, J., & Ochsner, K. N. (2010). Social cognitive conflict resolution: Contributions of domain-general and domain-specific neural systems. *Journal of Neuroscience* 30(25): 8481–8.
- Zaki, J., Lopez, G., & Mitchell, J. (2013). Person-invariant value: Orbitofrontal activity tracks revealed social preferences.
- Zaki, J., & Mitchell, J. (2011). Equitable decision making is associated with neural markers of subjective value. *Proceedings of the National Academy of Science, USA* 108(49): 19761–6.
- Zaki, J., & Ochsner, K. (2009). The need for a cognitive neuroscience of naturalistic social cognition. *Annals of the New York Academy of Science* 1167, 16–30.
- Zaki, J., & Ochsner, K. (2011). Reintegrating accuracy into the study of social cognition. *Psychology Inquiry* 22(3): 159–82.
- Zaki, J., & Ochsner, K. (2012). The neuroscience of empathy: Progress, pitfalls, and promise. *Nature Neuroscience* 15(5): 675–80.
- Zaki, J., Ochsner, K. N., Hanelin, J., Wager, T., & Mackey, S. C. (2007). Different circuits for different pain: Patterns of functional connectivity reveal distinct networks for processing pain in self and others. *Social Neuroscience* 2(3–4): 276–91.
- Zaki, J., Weber, J., Bolger, N., & Ochsner, K. (2009). The neural bases of empathic accuracy. *Proceedings of the National Academy of Science, USA* 106(27): 11382–7.

Mirror neuron system and social cognition

Christian Keysers, Marc Thioux, and Valeria Gazzola

For humans understanding and predicting the actions of others and being able to learn by observing the actions of a teacher are essential foundation for success. It is only in the last two decades, with the discovery of mirror neurons, that we start to have an understanding of how the brain enables these essential capacities. Even more recently, still, we start to understand that a similar mechanism may apply to how we perceive the sensations and emotions of others. Here, we will introduce the key methods used to study mirror neurons (directly with single cell recordings or indirectly using a range of non-invasive methods), review their properties, localization, and functions in monkeys and humans. Finally, we will discuss the possible extension of mirror neurons to sensations and emotions, and examine the psychiatric relevance of this system.

Mirror neurons in the monkey

Mirror neurons were first discovered in the ventral premotor cortex (area F5, Figure 14.1) of the macaque monkey (Fujii, Hihara, & Iriki, 2008; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). Neurons in region F5 are active, while the monkey executes goal-directed actions, such as grasping an object or shelling a peanut. For each premotor neuron, the execution of only a particular subset of all possible actions is linked to an increase in firing rate. This subset will be called ‘effective executed actions’, and represents a tuning curve of the premotor neuron. Electrostimulation in area F5 induce overt behavior such as grasping a nearby object (Graziano, Taylor, & Moore, 2002). This shows that neurons in F5 are part of the cascade of neurons that trigger a monkey’s own actions.

Unlike motor neurons in the primary motor cortex (M1) that code the nitty-gritty details of how the monkey will move his body, F5 neurons are akin to generals in the army: they determine what should be done, rather than precisely how it should be done (Rizzolatti, Camarda, Fogassi, Gentilucci, Luppino, & Matelli, 1988; Thioux, Gazzola, & Keysers, 2008). For instance, while different populations of M1 neurons control grasping an object with the hand and with the mouth, many F5 neurons respond similarly during these two actions. The firing of neurons in F5 is thus associated with a particular goal (grasping), rather than with a particular muscle movement (Rizzolatti et al., 1988).

Originally, F5 neurons were thought to deal exclusively with the execution of the monkey’s own actions. A significant number of F5 neurons, however, also responds while the monkey does not move its body, but simply views another individual perform certain actions. Actions, the observation of which trigger activity in a certain premotor neuron, will be called ‘effective observed actions’ (e.g. grasping or shelling a peanut), and form the visual tuning curve of premotor neurons. According to the relationship between the visual and motor tuning curve, visuo-motor F5 neurons can be classified as non-congruent (the tuning curves do not overlap) or as ‘mirror neurons’ (the

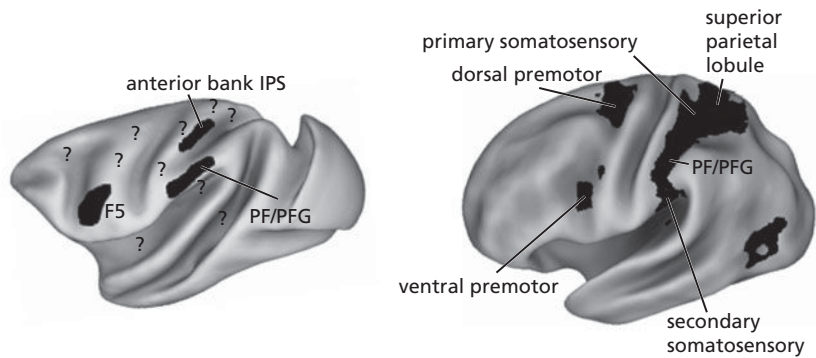


Figure 14.1 Localization of the mirror neuron system. Locations in which mirror neurons have been found in monkeys (left), with most of the brain still unexplored (question marks). Location of the putative MNS as identified using fMRI in humans.

Adapted from Gazzola, V. & Keysers, C. The observation and execution of actions share motor and somatosensory voxels in all tested subjects: single-subject analyses of unsmoothed fMRI data, *Cerebral Cortex* **19**, 1239–55. © 2009, Oxford University Press, with permission. For permission to reuse this material, please visit <http://www.oup.co.uk/academic/rights/permissions>.

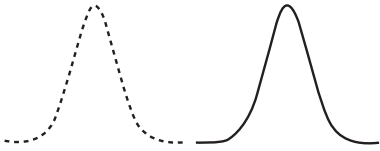
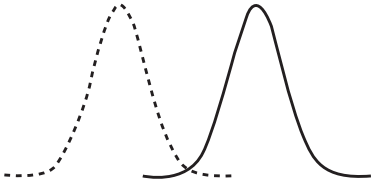
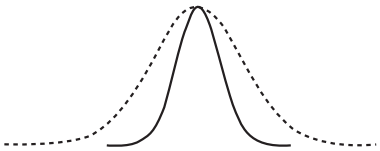
curves do overlap). Box 14.1 specifies the main classes of visuo-motor neurons encountered in F5 and indicates the proportion of neurons falling in each class.

Neurons with overlapping visual and motor tuning curves were dubbed ‘mirror neurons’ because when monkey A sees monkey B grasp a banana, for example, such neurons in A will correspond to the grasping motor programs of A. These grasping programs in A now mirror the activity in the brain of monkey B, which is currently also activating its grasping motor program, since he is grasping. The mirror neurons in A thus act like a mirror reflection or resonance of the motor programs that lead to the observed behavior in B. Non-congruent neurons should therefore not be called mirror neurons. The term ‘mirror neuron’ has been widely accepted in the literature, but it shouldn’t be taken to suggest that mirror neurons, like a true mirror, reproduce every detail of an observed action. Instead, mirror neurons produce something akin to an impressionist painting of the action that has been seen; in broad strokes, they capture the goal of the observed action through the observing monkey’s own, subjective motor vocabulary.

More recently, mirror neurons have also been recorded in the anterior half of the convexity of the inferior parietal lobule of the macaque (area PFG and to a lesser extent, PF (Rozzi, Ferrari, Bonini, Rizzolatti, & Fogassi, 2008), see Figure 14.1). These parietal mirror neurons have properties that are surprisingly similar to those of F5 neurons and are composed of similar proportions of non-congruent and mirror neurons. In both brain regions only 10–20% of neurons have mirror properties. A further set of mirror neurons have been reported in the anterior bank of the intraparietal sulcus (Fujii et al., 2008). The fact that mirror neurons have not yet been found in other locations in the macaque brain, however, cannot be taken as evidence that only F5, PF, PFG, and the intraparietal sulcus contain mirror neurons: mirror neurons have not been systematically searched for outside of these brain regions. Studies mapping the uptake of glucose in the macaque brain while monkeys observe and execute actions suggest that a much wider network of regions might be active during both conditions, including the somatosensory cortices (Evangelidou, Raos, Galletti, & Savaki, 2009; Raos, Evangelidou, & Savaki, 2004). Whether these regions, however, contain mirror neurons, or separate, but intermixed populations of neurons active during observation and execution remains to be understood.

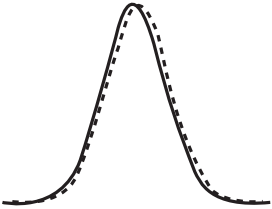
Box 14.1 Types of visuo-motor neurons

Neurons responding both during the execution and observation/audition of actions can be classified based on the relationship between their visual/auditory (dotted) and motor (solid) tuning curves. The last two columns of the table below indicate the approximate proportion of neurons in each category (as percentage of motor neurons responding to action observation) based on Gallese et al. (1996) for F5 and Rizzo et al. (2008) for PFG. Broadly and strictly congruent neurons are separated by a dotted line to indicate that, rather than a sharp distinction, one should see this difference as one of degree, with individual neurons varying along a continuum from very strict to very broad.

Family	Definition and example	Visual (dotted) and motor (solid) tuning curves	F5	PFG
Non-congruent	<p><i>Non-congruent</i>: there is no overlap between visual and motor tuning curves and no logical relationship</p> <p><i>Example</i>: effective executed action include tearing an object apart, but not grasping. Effective observed action include grasping, but not tearing.</p>		7%	11%
	<p><i>Logically related</i>: there is no overlap between visual and motor tuning curves, but a logical relationship</p> <p><i>Example</i>: effective executed actions include grasping, but not placing, while effective observed actions include placing, but not grasping. Given that an experimenter's placing food often precedes the monkey's grasping of that food, a logical relationship exists in the form of a causal chain of events</p>			6%
Mirror	<p><i>Broadly congruent</i>: there is overlap between motor and visual tuning curves, but the overlap is partial, with some actions only being effective in one of the modalities</p>		61%	54%

(Continued)

Box 14.1 (Continued)

Family	Definition and example	Visual (dotted) and motor (solid) tuning curves	F5	PFG
	<i>Example:</i> effective executed actions restricted to a precision grip with the hand, effective observed actions include precision grip with the hand, but also whole hand prehension and grasping with the mouth			
	<i>Strictly congruent:</i> the tuning curves overlap very tightly. <i>Example:</i> precision grip with the hand is the only effective action during both observation and execution		32%	29%

Region PF/PFG has reciprocal anatomical connections with the visual cortex lining the STS. Many neurons in the STS show visual responses that are similar to those of mirror neurons, but STS neurons lack the motor properties that make mirror neurons so unique (Keysers & Perrett, 2004; Matelli et al., 2006; Rozzi, Calzavara, Belmalih, Borra, Gregoriou; Rozzi et al., 2008). These STS neurons could send PF/PFG neurons visual information that causes the visual properties of mirror neurons in PF/PFG, and this information could then be transmitted from PF/PFG to F5 through the reciprocal connections between these regions (Keysers & Perrett, 2004; Rozzi et al., 2006). Recent neuroimaging studies using fMRI in monkeys have shown that the mirror neuron system may be composed of multiple circuits operating in parallel. One circuit transmits information from STS to PF/PFG and to the convexity of area F5, the other transmits information from the STS to the anterior intraparietal cortex to regions on F5 hidden in the arcuate sulcus. It remains unclear how these two routes differ, but the former may convey more information about the actions of the agent, and the latter about the properties of the object that is the goal of the action (Nelissen, Borra, Gerbella, Rozzi, Luppino, Vanduffel, et al., 2011).

Additional properties of monkey mirror neurons

Object-directed actions

Most of the mirror neurons in the convexity of F5 (F5c) respond exclusively to **object-directed** actions. A mirror neuron responding to the observation and execution of grasping a peanut for instance will not respond to seeing the same action mimed, i.e. performed without the object (Gallese et al., 1996; Umiltà, Kohler, Gallese, Fogassi, Fadiga, Keysers, et al., 2001). This matches the motor properties of F5c neurons, the vast majority of which seem to trigger actions directed at objects (i.e. ‘transitive’ actions). A minority of mirror neurons additionally responds to the sight of intransitive actions visually similar to the effective motor and visual action: a neuron responding

during grasping an object with the mouth might also respond to the sight of lip-smacking, an affiliative gesture frequently performed by monkeys (Ferrari, Gallese, Rizzolatti, & Fogassi, 2003). These neurons are thought to serve as an important pre-adaptation for the passage from actions to symbolic gestures, and thereby for the emergence of language (Keysers, 2011). The preference for transitive actions observed in F5c also holds for parietal area PFG, but it remains unclear whether other brain regions may contain a higher proportion of mirror neurons preferring intransitive actions (e.g. dorsal premotor cortex).

Occlusion

About 50% of the mirror neurons in F5c responding to the sight of grasping an object, but not to the sight of a hand miming a grasp, also respond to seeing a hand reaching behind an occluding screen, if and only if the monkey has previously seen an object being hidden behind that screen (Umiltà et al., 2001). Mirror neurons therefore seem part of a circuitry combining present, but incomplete visual information with past information stored in memory.

Sound

Some mirror neurons also respond to the sound of their effective executed actions. Many mirror neurons responding during the execution and observation of peanut shelling also respond to the sound of peanut shelling. Some only respond when the sound and the vision of the effective action occur contemporarily, others respond to either modality alone, but more to their combination, and others will respond maximally to either modality (Keysers, Kohler, Umiltà, Nanetti, Fogassi, & Gallese, 2003; Kohler, Keysers, Umiltà, Fogassi, Gallese, & Rizzolatti, 2002). It is often difficult to determine the latency of visual mirror neurons, because the sight of most actions unfolds over hundreds of milliseconds (e.g. reaching, grasping then shelling a peanut), and it then becomes difficult to determine whether a given spike is a late response to an early component of the action (e.g. reaching), or a fast response to a later component (e.g. shelling), given that early components could be predictors for later components. In the auditory modality, this is easier to do, as only the final interaction with the object (e.g. shelling) produces the sound. On average, we found auditory mirror neuron to start responding 120ms after the onset of the sound of the effective action, suggesting a rapid and relatively direct route from the sensory input to mirror neurons (Keysers et al., 2003; Kohler et al., 2002).

Action discrimination

It is possible to deduce which of two actions another individual is performing with >90% accuracy from the firing rate of F5c mirror neurons in the brain of an observing and/or listening monkey (Keysers et al., 2003), which is not lower than the accuracy during motor execution. That the firing of premotor neurons carries so much information about the actions of other individuals is remarkable given that a decade ago, the premotor cortex was considered not to have any functions in perception.

Embedding in larger action schemes

Specific motor acts (e.g. grasping) can be embedded into larger action sequences (e.g. grasping to eat or grasping to place). About half of F5 and PFG neurons change their firing rate during the execution of a specific act based on what sequences the act is embedded in (Bonini, Serventi, Simone, Rozzi, Ferrari, & Fogassi, 2011). Interestingly, this is true for mirror neurons in both regions as well—a grasping neuron that responds more to grasping to eat than grasping to place will also respond more to the sight of someone else grasping to eat than grasping to place, and vice versa (Bonini, Rozzi, Serventi, Simone, Ferrari, & Fogassi, 2010). Obviously, the neurons

cannot magically know, whether someone else will grasp to eat or place, but in these experiments the type of object being grasped (food vs. non-edible object) and the presence of a cup into which the object needs to be placed helped the monkey know what context the act is embedded into. If the monkey is prevented from witnessing such cues, the selectivity disappears (Bonini et al., 2011).

The how, what and why of actions

If we sit together at a bar, and I see you take some peanuts, this action has three levels of description: **How**, e.g. you are cupping the peanuts with your hand, you are precision grasping a single peanut, or you are using a spoon; **What**, e.g. you are taking the peanuts or your glass and **Why**, e.g. you are grasping them to eat them yourselves, or to hand them over to me. The distinction between strictly congruent mirror neurons and broadly congruent mirror neurons draws a distinction between how and what. Because strictly congruent mirror neurons care about the exact way in which actions are performed, they can provide information about **how**. Because broadly congruent mirror neurons often respond similarly to actions achieving the same purpose with different means, they provide information about **what** is being performed (e.g. grasping) independently of how (e.g. with the hand or mouth). Finally, the selectivity of ~50% of mirror neurons in PFG and F5c for the sequence in which an action is embedded suggests that these neurons additionally contain information about the **why** of the action. Thus, because F5 and PFG contain a mix of these different mirror neurons (strictly and broadly congruent, with and without sequence preference), these regions contain information about all three levels of description. With regard to why, it is likely however, that the highest level of why remains opaque to these regions. It is unlikely, for instance, that response properties in these regions could specifically encode that you hand me these peanuts to see the surprise and pain the wasabi coating will trigger on my face (Thioux et al., 2008). Thinking about the reason why someone is doing something generates an increase of activity in mentalizing brain regions known to be involved in processing the state of mind of other people (Spunt, Satpute, & Lieberman, 2011), which indicates that at least some of the inferences concerning the why of an action are processed outside the mirror neuron system.

The human mirror neuron system

The definition of mirror neurons includes that a single neuron be involved both during action execution and during the perception (observation/listening) of the same action. In humans, this definition is challenging, because the activity of single neurons can only rarely be recorded (Keysers & Gazzola, 2010; Mukamel, Ekstrom, Kaplan, Iacoboni, & Fried, 2010). However, the existence of a system that has the same properties as the mirror neurons in monkeys, and which is therefore called the putative mirror neuron system (pMNS) is now well established in humans by more than a decade of experiments based on a variety of less invasive techniques (see Box 14.2). Taken together, these techniques have established two basic facts.

Existence

Measuring TMS evoked motor potentials (Avenanti, Bolognini, Maravita, & Aglioti, 2007; Aziz-Zadeh, Iacoboni, Zaidel, Wilson, & Mazziotta, 2004; Fadiga, Craighero, & Olivier, 2005; Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995; Urgesi Maieron, Avenanti, Tidoni, Fabbro, & Aglioti, 2010) and motor reaction times (Brass, Bekkering, Wohlschlaeger, & Prinz, 2000), while participants perceive the actions of others has shown that viewing or hearing actions facilitates the execution of corresponding actions and interferes with the execution of antagonistic actions. This can only

Box 14.2 How to study the mirror neuron system in human

Based on the definition of mirror neurons in the monkey, a brain area should only be considered to be putatively mirror if it is involved **both** in action observation (or listening) **and** execution. A number of techniques have been used to establish whether this is true of some neurons and/or brain regions in humans.

TMS (transcranial magnetic stimulation)

A single magnetic pulse, given on a particular location of M1's homunculus, evokes a twitch in a corresponding muscle that is measured using electromyography, leading to a MEP measurement. If a human MNS exists, seeing or hearing someone else perform an action involving the same muscle should increase the MEP, but one involving other muscles should not (see, for example, Fadiga et al., 1995). Repetitive TMS (rTMS) can also be used to interfere with normal processing in a restricted area of the brain shortly before using the MEP technique described above to test if this particular area is part of the MNS that causes the MEP modulation during observation (for example Avenanti et al., 2007).

Psychophysics

Instead of triggering motor programs using TMS one can also ask participants to execute certain actions in response to a cue and simultaneously show seemingly task irrelevant actions. Again, if there is a MNS, the execution of a particular action should be *accelerated* by viewing actions using the same muscle and should be *slowed* by viewing actions using an antagonistic muscle (for example, see Brass et al., 2000; Kilner et al., 2003).

PET (positron emission tomography) and fMRI (functional magnetic resonance imaging)

The distribution of injected radioactive molecules of water or oxygen (PET), or the distribution of changes in the distortions of the magnetic field due to blood oxygenation (fMRI) can be used to localize brain regions activated while viewing or hearing the actions of others. The same procedure is then used to visualize the brain regions involved in action execution, preferably in the same participants. Only voxels activated both during the perception and execution of actions will be considered part of the putative MNS (for example, Gazzola & Keysers, 2009).

Repetition suppression fMRI

For many neurons, the repetition of the same stimulus causes a reduction of the activity to that stimulus. If action observation and execution are indeed encoded by the same (population of) neurons, asking the subject to watch the very same action he/she just performed (or the other way around) should cause a reduction of the activity similar to the one caused by the repetition of two consecutive visual stimuli. Areas showing this reduction will then be considered part of the putative MNS (for example, Kilner et al., 2009). While this method initially encountered much enthusiasm, it is now rather contested (Bartels et al., 2008). This is because if repetition suppression is found in a region, neurons in that region may still not encode the property that

(Continued)

Box 14.2 (Continued)

triggered the suppression (Bartels et al., 2008). Conversely, if repetition suppression does not occur in a particular region, this may be due to neurons in this region not showing repetition suppression or to the region not containing enough mirror neurons to trigger a measurable suppression or that repetition suppression mislocates the relevant neurons. In our own studies, mirror neurons in monkeys indeed failed to show repetition suppression (Keysers et al., 2003).

Pattern classification fMRI

If a MNS exists, the pattern of activity across pMNS regions should be similar during action perception and execution. Accordingly, after training a pattern classifier to discriminate brain activity when participants hear or view actions A and B, the algorithm should automatically discriminate above chance the patterns associated with executing actions A and B (for example, Etzel et al., 2008).

EEG/MEG (electroencephalogram or magnetoencephalogram)

EEG and MEG measure, through the scalp, currents that are generated by synchronous activity of populations of neurons. The mu-rhythm amongst that signal has more energy while the subject is at rest and is disrupted as soon as an action is performed. If a MNS exists a similar suppression should occur while perceiving the actions of others. An alternative approach is to stimulate the median nerve, which produces a disruption of the mu and beta rhythms similar to the one produced by action execution. Studying whether this induced disruption is modified by action observation is another method to assess the existence of a MNS in humans (for a review, see Pineda, 2005). Finally, using source localization, one can test if activity during action perception depends on sources that overlap with those during action execution, but the limited spatial resolution of source localization makes this approach less attractive.

be the case if at least some neurons involved in executing a motor program receive excitatory input from sensory neurons responding to the sight or sound of the same action. Interestingly, repetitive TMS over the premotor cortex, but not the primary motor cortex reduces this facilitation, indicating that this convergence of sensory and motor information occurs in the premotor cortex, where mirror neurons exist in monkeys (Avenanti et al., 2007). These findings are highly compatible with the notion of a human mirror neuron system, as mirror neurons need to get sensory information regarding their preferred action. However, mirror neurons in the strict sense need to increase their firing rate during both the execution and the perception of an action. Whether that is the case, cannot directly be deduced from these experiments—synaptic facilitation can be strong enough to modulate the activation of a motor neuron while the individual is performing an action, but too weak to trigger spiking at rest. Additionally, Avenanti et al. also noticed that disturbing the activity of SI can lead to a reduction of the visual facilitation of action execution if the visual stimulus contains salient somatosensory components (e.g. joint stretching). This suggests that the somatosensory system might play a role in the pMNS.

Independent evidence for a recruitment of the motor and somatosensory system during action observation stems from the observation that perceiving the actions of others is linked with changes

in two frequency bands of the power-spectrum of the EEG and MEG signal that resemble those associated with executing similar actions. These two bands are jointly called the mu-rhythm, with the lower frequency band around 10 Hz being referred to as alpha or lower alpha, and the higher 20 Hz band referred to as beta or upper alpha. This visual modulation of a motor rhythm was first observed in 1954, four decades before the discovery of mirror neurons, by Gastaut and Bert in a surprisingly modern experiment: “[the rolandic mu-rhythm] is blocked when the subject performs a movement ... It also disappears when the subject identifies himself with an active person represented on the screen ... During a sequence of film showing a boxing match ... less than a second after the appearance of the boxers all type of rolandic activity disappears in spite of the fact that the subject seems completely relaxed” (Gastaut & Bert, 1954, p. 439). Once mirror neurons were described, this phenomenon received renewed interest, with a number of experiments now confirming its existence using EEG (Cochin, Barthelemy, Lejeune, Roux, & Martineau, 1998; Cochin, Barthelemy, Roux, & Martineau, 1999; Muthukumaraswamy & Johnson, 2004b; Muthukumaraswamy & Johnson, 2004a; Muthukumaraswamy et al., 2004; Pineda, 2005) and MEG (Hari, Forss, Avikainen, Kirveskari, Salenius, & Rizzolatti, 1998).

Most recently, a small number of neurons have been tested for mirror properties in the human cortex (Keysers & Gazzola, 2010; Mukamel, et al., 2010). This was done while patients had electrodes implanted to localize the origin of epileptic seizures. These recordings evidenced 11 neurons that behaved like broadly congruent mirror neurons in the monkey, with some responding during the execution and observation of hand actions (whole hand or precision grasp), and others during the execution and observation of facial expressions (smile or frown). Because the localization of the electrodes is dictated by the likely source of the seizures, they did not include ventral premotor cortex or inferior parietal regions, but instead, the supplementary motor area and the medial temporal lobe, both of which turned out to contain neurons responding during action observation and execution.

Localization

Secondly, PET and fMRI studies that map brain activity in the same participants during (i) action execution and (ii) viewing or hearing others perform similar actions have helped map regions that could contain mirror neurons in humans. A number of brain regions were found to respond with augmented blood flow or BOLD signal in both cases (Figure 14.1, see (Keysers, Kaas, & Gazzola, 2010; Caspers, Zilles, Laird, & Eickhoff, 2010; Keysers & Gazzola, 2009) for reviews). This network includes ventral premotor and inferior parietal cortices, which contain mirror neurons in monkeys, and additional areas including the dorsal premotor, supplementary motor, primary and secondary somatosensory, dorsal posterior parietal cortex and the cerebellum (Aziz-Zadeh, Wilson, Rizzolatti, & Iacoboni, 2006; Caspers et al., 2010; Dinstein, Hasson, Rubin, & Heeger, 2007; Gazzola, Aziz-zadeh, & Keysers, 2006; Filimon, Nelson, Hagler, & Sereno, 2007; Gazzola, Rizzolatti, Wicker, & Keysers, 2007a; Grezes, Armony, Rowe, & Passingham, 2003; Keysers & Gazzola, 2009; Ricciardi, Bonino, Sani, Vecchi, Guazzelli, Haxby et al., 2009; Turella, Erb, Grodd, & Castiello, 2009). However, the limited spatial resolution of traditional fMRI and PET cannot show that the same neurons within a voxel are responsible for the augmented BOLD response during action execution and action perception, so it is therefore appropriate to refer to this network as the “putative MNS,” where putative acknowledges the fact that in some of these regions, different neurons within a voxel could be responsible for the activity during action execution and action perception. Additionally, BOLD fMRI is known to be sensitive to synaptic activity even in the absence of changes in firing rate of the neurons (Lippert, Steudel, Ohl, Logothetis, & Kayser, 2010) and augmentations in BOLD during action observation and execution could reflect subthreshold synaptic

inputs, rather than neural firing in some of the cases. Putative however does not question the existence of a MNS in humans *per se*, because the existence is now established by the single cell recordings (Mukamel et al., 2010).

Repetition suppression fMRI (see Box 14.2) has been used to provide further evidence that at least the human premotor and posterior parietal cortex also contain mirror neurons (Chong, Cunnington, Williams, Kanwisher, & Mattingley, 2008; Dinstein et al., 2007; Kilner, Neal, Weiskopf, Friston, & Frith, 2009; Lingnau, Gesierich, & Caramazza, 2009). However, repetition suppression has been shown to mislocalize certain neural properties (Bartels, Logothetis, & Moutoussis, 2008), and results derived from that method should be considered with great care. Also, if Blood Oxygen Level Dependent (BOLD) mainly measures synaptic processes, repetition suppression is likely only to work if the sensory and motor input to mirror neuron comes through the same synapses, which is not necessarily the case.

Although mu-suppression has been widely used to show that the motor system is recruited during action observation, until recently, the limited resolution of EEG made it difficult to localize what brain regions were responsible for mu-suppression during action observation and execution. In a recent study (Arnstein, Cui, Keysers, Maurits, & Gazzola, 2011), EEG was measured inside the scanner while simultaneously measuring fMRI BOLD. Participants were shown movies of hand actions and control stimuli, and were asked to perform simple motor actions in the scanner. Results indicate that trials with high mu-suppression were trials in which BOLD somatosensory and dorsal premotor activity was high during both action observation and execution, but not trials in which the ventral premotor activation was high. Thus mu-suppression EEG and classical fMRI during action observation do measure overlapping processes, but mu-suppression may not originate from the ventral premotor cortex that has historically been most associated with the MNS, but with a number of regions that have received less attention, including the somatosensory cortex.

Finally, with the advent of pattern classification methods to look at the distributed pattern of activation in regions of the cortex during action observation and execution, it has become possible to examine whether the pattern of activity is indeed similar while participants perform and perceive actions. This is important, because neuroscience converges to attribute representations to distributed patterns of activity of many neurons, rather than with single neurons. In the first successful study of that kind, it was shown that premotor, posterior parietal and somatosensory cortices indeed show similar patterns of activation while performing and listening to actions (Etzel, Gazzola, & Keysers, 2008). Shortly thereafter, similar results were obtained for performing and viewing actions (Oosterhof, Wiggett, Diedrichsen, Tipper, & Downing, 2010).

Properties of the human mirror neuron system

Object and non-object directed actions

In humans, there is evidence that the pMNS also responds to actions not directed at objects (intransitive actions). EEG/MEG studies show a suppression of the mu-rhythm during finger movements observation (Babiloni, Babiloni, Carducci, Cincotti, Coccozza, Del Percio, et al., 2002; Calmels, Holmes, Jarry, Hars, Lopez, Paillard, et al., 2006; Cochin et al., 1998, 1999); TMS studies show a motor-evoked potential (MEP) facilitation during the observation of finger flexion and extension (Baldissera, Cavallari, Craighero, & Fadiga, 2001; Clark, Tremblay, & Ste-Marie, 2004; Fadiga et al., 1995); fMRI studies find premotor activity while observing finger lifting (Iacoboni, Woods, Brass, Bekkering, Mazziotta, & Rizzolatti, 1999), hand gestures (Schippers, Gazzola, Goebel, & Keysers, 2009; Schippers, Roebroek, Renken, Nanetti, & Keysers, 2010; Lui, Buccino, Duzzi, Benuzzi, Crisi,

Baraldi, et al., 2008) and the pantomime of object related actions without the object (Buccino, Binkofski, Fink, Fadiga, Fogassi, Gallese, et al., 2001); and psychophysical studies show that seeing another person moving their fingers (Brass et al., 2000) or arm (Kilner, Paulignan, & Blakemore, 2003) facilitates the execution of similar movement. Nevertheless when intransitive and transitive actions are directly compared, transitive actions seem to elicit more pMNS activity (Buccino et al., 2001; Muthukumaraswamy et al., 2004), but see (Rossi, Tecchio, Pasqualetti, Ulivelli, Pizzella, Romani, et al., 2002).

How, What and Why

In the monkey, the combination of strictly and broadly congruent mirror neurons enables the MNS to contain information about **what** was performed (the goal) and **how** it was performed (the means). In humans, too, there is evidence for the representation of both goals and means in the pMNS (Thioux et al., 2008). TMS experiments evidence selective facilitation of the muscles involved in the observed action (Alaerts, Heremans, Swinnen, & Wenderoth, 2009; Urgesi, Candidi, Fabbro, Romani, & Aglioti, 2006). fMRI experiments show that actions with familiar goals, but means the actor cannot reproduce still triggers activation in regions of the motor cortices recruited while the observer performs actions within its motor program that have the same goal (Aziz-Zadeh, Sheng, Liew, & Damasio, 2011; Gazzola et al., 2007a; Gazzola, Van Der Worp, Mulder, Wicker, Rizzolatti, & Keysers, 2007b), possibly allowing us to mirror at least the goal of non-human (robotic or animal) actions as well. Additionally, repetition suppression fMRI should be interpreted with care (Box 14.2), but it has been found that parts of the pMNS respond less to the sight of an action if it is preceded by an action sharing the same goal (Hamilton & Grafton, 2006, 2008). Because in monkeys, broadly congruent mirror neurons are twice as numerous as strictly congruent ones (Gallese et al., 1996; Rozzi et al., 2008), goals should prevail over means in the MNS, a finding compatible with the observation that if asked to reproduce an action, humans tend to reproduce the goal rather than the means (Bekkering, Wohlschläger, & Gattis, 2000), but specific task instructions can modify that tendency toward means (Bird, Brindley, Leighton, & Heyes, 2007a). Finally, in analogy to the sequence selectivity in humans, the human pMNS seems to be differentially activated by grasping in contexts that suggest different future actions (Iacoboni, Molnar-Szakacs, Gallese, Buccino, Mazziotta, & Rizzolatti, 2005). This suggests a certain level of ‘why’ coding in the pMNS. More abstract aspects of why however most likely require additional brain regions (Schippers et al., 2010; Spunt et al., 2011; Thioux et al., 2008).

Sounds

The human pMNS, as its macaque equivalent, is also activated by the sound of an action (Caetano, Jousmaki, & Hari, 2007; Gazzola et al., 2006; Pizzamiglio, Aprile, Spitoni, Pitzalis, Bates, D’Amico, et al., 2005), even in congenitally blind participants (Ricciardi et al., 2009) showing that the pMNS can develop independently of vision. In terms of what vs. how, auditory responses in the pMNS are a particularly interesting case. While the sight of an action usually involves information about the means and the goal (we see a hand press a piano key), the sound of an action often carries no information about what body part was used to perform the action (e.g. one could press a piano key with the foot, nose or hand and produce the same sound). Auditory mirror responses therefore epitomize the pMNS’s capacity to mirror what was done (press a key) without direct information about how it was done, probably by activating the motor programs the listener would use most frequently to produce the same sound. It shows that the MNS projects the perceiver’s own

motor programs onto the sensory evidence of other people's actions, rather than objectively mirroring the details of how the other has performed the action. Interestingly, the response of auditory pMNS has been shown to be plastic into adult life (see "Plasticity and development").

Correlation with empathy

People that report being more empathic in life, as measured using self report empathy questionnaires (IRI (Davis, 1980), or EQ (Baron-Cohen & Wheelwright, 2004)) have sometimes been found to activate region associated with the MNS more strongly (see Baird, Scheffer, & Wilson, 2011 for a critical review). Positive correlations were found between premotor and SI activation to the sound of neutral hand actions and the perspective taking subscale of the IRI (Gazzola et al., 2006). Similarly, during the observation of facial expressions, higher ventral premotor activation was measured in participants with higher scores on the emotional subscales of the IRI (Jabbi, Swart, & Keysers, 2007), on the IRI total score (Pfeifer, Iacoboni, Mazziotta, & Dapretto, 2008) or on the EQ (Chakrabarti, Bullmore, & Baron-Cohen, 2006). Lesions to the ventral premotor cortex have also been found to reduce self reported empathy in the emotional subscales, confirming an association of this region with empathy (Shamay-Tsoory, Aharon-Peretz, & Perry, 2009). The association of self-reported empathy and pMNS activation is, however, not extremely robust. In the same sample in which we found a correlation between pMNS activation and the perspective taking score while listening to the sound of actions, we failed to find a similar correlation while viewing the actions of others. In a number of other studies in our lab, we also failed to find similar associations. Because studies typically publish correlations if they find them, but not if they do not, it is difficult to judge what proportion of studies that correlated IRI scores with pMNS activation indeed found such correlation. The key question may be under what conditions, differences in self-report empathy lead to measurable differences in pMNS activations.

Anticipation

While observing the actions of others in a predictable context, activity in the MNS seems to predict the actions of others by ~200 ms. Infants suppress their mu-rhythm a couple of hundreds of milliseconds before the onset of a predictable arm movement (Southgate, Johnson, Osborne, & Csibra, 2009). Viewing a hand rhythmically flex and extend the wrist, adult observers will modulate the excitability of their own wrist muscles with the same frequency, but about 200ms in advance of the observed movement (Borroni, Montagna, Cerri, & Baldissera, 2005). TMS-induced motor evoked potentials are facilitated by the vision of a grasping action, but more so when viewing a frame taken before the grasp than by a frame depicting the grasping itself (Urgesi et al., 2010). Given that it takes at least 100 ms for the brain to relay auditory and visual information to the MNS (Keysers et al., 2003; Kohler et al., 2002), this suggests that the MNS is part of a system that will anticipate predictable events. Interestingly, while exploring the flow of information in the MNS while viewing the actions of others, we found that although the information flow is initially from visual to premotor regions, the flow later inverts, with a preponderance of backward flow of information from premotor to visual regions (Schippers & Keysers, 2011). This observation is compatible with the notion that the MNS serves to predict upcoming sensory evidence about the actions of others (Gazzola & Keysers, 2009; Schippers & Keysers, 2011). Such anticipation may become particularly important when two agents have to act jointly: our sensorimotor delays normally make us react with a delay of ~200ms to sensory signals, anticipating the actions of others by about 200ms would then ensure that we can act in real time to the (anticipated) actions of others (Kokal & Keysers, 2010).

Plasticity and development

Expertise

fMRI experiments suggest that expertise in a domain is associated with increased activity in the putative MNS to the sight/sound of actions of that domain. Ballet dancers who have learned a specific dance style have higher level of activity in the pre-motor and parietal cortex while watching that dance relative to a dance unknown to them (Calvo-Merino, Glaser, Grezes, Passingham, & Haggard, 2005). Professional pianists activate the putative MNS more than piano novices when viewing a hand play the piano (Haslinger, Erhard, Altenmüller, Schroeder, Boecker, & Ceballos-Baumann, 2005).

Training

Practice-related changes in the activation of the putative MNS can be observed after relatively short periods of training. Five weeks of dance training boosts activations in the pMNS when observing the learned dance sequences relative to other matched sequences (Cross, Hamilton, & Grafton, 2006) and this increase correlated with the dancers' appreciation of their ability to reproduce the sequences. Also, although piano naïve participants do not significantly activate their premotor cortex to the sound of piano music, after only five half-hour lessons of piano, they do (Lahav, Saltzman, & Schlaug, 2007). Finally, when participants first observe movies involving index and little-finger movements, MEPs increase in the index, but not the little finger for the observation of index finger abduction, and vice-versa for the observation of little-finger abduction. After being trained to perform index abductions when seeing little-finger movements and vice-versa for a couple of hours, participants in this group showed a reversed effect: facilitation of MEPs of the index abductor when viewing little finger abduction and vice versa (Catmur, Walsh, & Heyes, 2007).

Hebbian learning

Since an actor is spectator and auditor of her own actions, during hand actions for instance, parietal and pre-motor neurons controlling an action fire at the same time as neurons in the STS that respond to the observation and sound of this specific hand action (some of which irrespective of the view point). These sensory and motor neurons that fire together would wire together, i.e. strengthen their connexions through Hebbian synaptic potentiation (Keysers & Perrett, 2004) (see also Heyes (2001) for a similar model based on association learning). After repeated self-observation/audition, the motor neurons in the premotor and parietal regions would now receive such strong synaptic input from sensory STS neurons responding to the sight and sound of the action, that they would become mirror. The same pairing between execution and observation would also occur in cases in which an individual is imitated by another (Brass & Heyes, 2005; Heyes, 2001; Del Giudice, Manera, & Keysers, 2009). For instance, a child cannot observe its own facial expressions, but the adult who imitates the child's expression would serve as a mirror, triggering in the child's STS an activity pattern, representing what the expression sounds and looks like, that becomes associated with the pre-motor cortex activity producing the expression that was imitated (Del Giudice et al., 2009). Hebbian learning could explain the emergence of the MNS in infants and its plasticity in adulthood. This perspective does not preclude the possibility that some genetic factors may guide its development. Genetic factors could for instance canalize (Del Giudice et al., 2009) Hebbian learning by equipping the baby with a tendency to perform spontaneous and cyclic movements and to look preferentially at biological motion congruent with its actions to provide the right activity patterns for Hebbian learning to occur. Also, it does not preclude that

mirror neurons may exist at birth for certain actions (e.g. tongue protrusion, see below). What is important in this perspective is that the MNS is no longer a specific social adaptation, that evolved to permit action understanding, but is a simple consequence of sensory-motor learning that has to occur for an individual to be able to visually control his own actions (Brass and Heyes, 2005; Del Giudice et al., 2009; Oztot & Arbib, 2002). Note that due to sensorimotor latencies, there is a systematic time-lag between motor activity and sensory consequences that endow this Hebbian learning with predictive properties (see “Functions of the motor mirror neuron system”). Hebbian learning would of course work both ways: not only would it train connections from sensory to motor cortices, but also the reverse connections, from motor to sensory cortices. The latter could then serve to anticipate what the actor should see in ~200 ms, based on the motor program it currently executes, and these internal models could explain why information from premotor to sensory cortices is observed during action observation and action execution (Gazzola & Keysers, 2009; Schippers & Keysers, 2011).

Neonatal imitation

Imitative abilities in newborns are often taken to suggest that the MNS is partially genetically predetermined. Infants younger than a month tend to imitate some facial expressions (Meltzoff & Moore, 1977). Tongue protrusion is the most frequently imitated behavior, but some experiments also report lips protrusion, mouth opening, eye-blinking, and finger movements imitation (Anisfeld, 1991; Meltzoff & Decety, 2003; Meltzoff & Moore, 1977). What remains unclear, is how this neonatal imitation relates to the MNS (Brass & Heyes, 2005; Heyes, 2001; Meltzoff & Decety, 2003; Meltzoff & Moore, 1977). It could represent the activity of a primitive and specialized system that is fundamentally different from the adult MNS or a genetic pre-wiring of what will become the adult MNS. Longitudinal studies that measure the neural basis of such neonatal imitation and follow it until mature imitation arises later in life will be necessary to settle this debate.

MNS activity in children

Few studies have measured putative MNS activity in children. EEG shows mu-suppression during action observation in children aged 4 to 10 years with the degree of mu-suppression independent of age (Lepage & Theoret, 2006). In addition, babies aged 14–16 months show more mu- and beta-suppression, while they view other babies crawl (which they can do themselves) than when they see other babies walk (which they cannot yet do) (van Elk, Van Schie, Hunnius, Vesper, & Bekkering, 2008). This effect was stronger in more experienced walkers. The change in EEG rhythms occurring during the first years of life however makes it difficult to extend this EEG approach to infants below the age of one year. Shimada and Hiraki (Shimada & Hiraki, 2006) therefore used near infrared spectroscopy, to test if younger children already activate their motor cortices while viewing the actions of others. They found that even 6–7-month-old infants already activate brain regions involved in hand action execution, while they view the hand actions of others. This suggests that action perception and execution may already be coupled by a MNS in 6-month-old babies, and that this system changes relatively little from 4 to 10 years of age.

Action understanding in infants

By the end of the first year, infants readily recognize the goals of observed actions. For instance, 6-month-old infants were habituated to seeing an experimenter grasp one of two toys (Woodward, 1998). After habituation, the position of the toys was inverted. Infants suddenly looked longer when the experimenter started picking the other toy (despite the fact that the movement trajectory

was the same as during habituation). They were less interested when the experimenter picked the old toy in the new location, suggesting that infants at that age already encode the goal of an observed action (i.e. what was grasped). This effect depends on the infant's ability to grasp. Three-month-old babies cannot yet reach and grasp objects by themselves, and they do not show the above-mentioned goal habituation. If they are fitted with a 'sticky' glove to which toys adhere, and become able to retrieve toys by themselves, when re-tested in the observation condition they start to show the goal habituation effect (Sommerville, Woodward, & Needham, 2005). The results of these habituation studies were recently corroborated by the analysis of anticipative eye-gaze behaviors in infants between 4 and 10 months of age (Kanakogi & Itakura, 2011). The experiment demonstrates that infants' ability to predict the goal of a grasping action (by gazing at the target in advance of the grasp) develops with their ability to perform the same action. This link between the capacity to perform an action (grasping) and the capacity to understand similar actions in others suggests that the MNS, linking observation and execution, might be involved.

Similar systems for emotions and sensations

Touch

Similarly to what occurs in the MNS for actions, in an fMRI experiment we showed that the secondary, and to a lesser extent, primary somatosensory cortex is activated not only when participants are touched on their own body, but also when seeing others be touched in similar ways (Keysers, Wicker, Gazzola, Anton, Fogassi, & Gallese, 2004). A number of fMRI studies have now confirmed this original finding (Blakemore, Bristow, Bird, Frith, & Ward, 2005; Ebisch, Perrucci, Ferretti, Del Gratta, Romani, & Gallese, 2008). In addition, Brodmann Area 2, a sector of the primary somatosensory cortex responsible for combining tactile and proprioceptive information, becomes active both when participants manipulate objects and when they see or hear other's do so (Keysers et al., 2010). This suggests that a mechanism similar to mirror neurons in the motor domain may apply to the somatosensory domain—we activate brain regions involved in our own tactile and proprioceptive experiences when seeing those of others. A recent single cell recording study in monkeys has confirmed that some neurons with somatosensory properties indeed also respond to the sight of other people being touched in parietal area VIP (Ishida, Nakajima, Inase, & Murata, 2010).

The case of mirror-touch synaesthesia is probably an instance of unusually high vicarious activation of this neuronal circuit for touch. A little over 1% of people experience touch upon seeing someone else being touched (Banissy, Cohen Kadosh, Maus, Walsh, & Ward, 2009). Touch is typically experienced on the same body part that was seen to be touched. One such individual was scanned using fMRI while observing videos of someone being touched (Blakemore et al., 2005). The study revealed a hyper-activity in the somatosensory cortices, the pre-motor cortex, and the anterior insula of this person relative to control participants. The phenomenon seems therefore to correspond to the over-activity of one component of the mirror neuron system, where the re-enactment of the feeling of touch is operated. Interestingly, individuals presenting this form of synaesthesia seem to be particularly empathic (Banissy & Ward, 2007) and show superior performance in the recognition of facial expressions (Banissy, Garrido, Kusnir, Duchaine, Walsh, & Ward, 2011), further strengthening the idea that empathy—the capacity to feel what goes on in others—depends on vicarious activation of similar states in the self.

Pain

Functional MRI and EEG studies suggest that witnessing the pain of other individuals activates regions of the pain matrix involved in feeling pain, including the somatosensory, insular and

anterior cingulate cortices (for reviews, see Lamm, Decety, & Singer, 2011; Keysers et al., 2010). Rare single cell recordings in epileptic patients suggest that single neurons in the anterior cingulate cortex indeed respond to both the experience of pinpricks and the observation of other individuals being pinpricked (Hutchison, Davis, Lozano, Tasker, & Dostrovsky, 1999). In addition, seeing a needle penetrate someone else's skin leads to relaxation of the muscles in the observer that correspond to the pricked region, suggesting an interaction between the mirroring of pain and the motor system (Avenanti, Minio-Paluello, Bufalari, & Aglioti, 2009). These effects are stronger in more empathic individuals (Avenanti et al., 2009; Singer Seymour, O'Doherty, Kaube, Dolan, & Frith, 2004). Patients with congenital insensitivity to pain, caused by a lack of peripheral nociceptors, allow investigating whether first-hand experience of **somatic** pain is necessary for empathizing with the pain of others. Compared to control participants, these patients report normal ratings of other people's pain if they can see the facial expressions, but not if this information is removed (Danziger, Prkachin, & Willer, 2006). These patients also show relatively normal, although less consistent, vicarious activations in the insula and anterior cingulate cortex when witnessing the pain of others (Danziger, Faillenot, & Peyron, 2009). Somatic pain is however not the only form of pain—pain experience and pain matrix activation along with pain facial expressions can be triggered by social factors including the separation from a loved one or social exclusion (Eisenberger, 2012). Because this social pain does not require the small diameter nociceptors these patients are lacking, congenitally pain insensitive patients might vicariously leverage social pain experiences to empathize with the somatic and social pains of others. Whether participants that never experienced any form of pain are capable of empathy for pain thus remains unknown.

Emotions

Finally, when observing the emotional facial expressions of others, participants not only activate regions they would use to generate similar expressions (van der Gaag, Minderaa, & Keysers, 2007)—this motor sharing appears to trigger activity in the anterior insula of the observer in regions that become active when the observer experiences similar emotions (Jabbi & Keysers, 2008; Jabbi et al., 2007; Wicker, Keysers, Plailly, Royet, Gallese, & Rizzolatti, 2003, for a review see Bastiaansen, Thioux, & Keysers, 2009). Again, this phenomenon is more pronounced in more empathic individuals (Jabbi et al., 2007).

The vicarious brain

Together, this data suggests that mirror neurons in the motor system may be a specific example of a more general phenomenon: our brain vicariously activates representations of our own actions, sensations and emotions while viewing those of others (Keysers, 2011). Lesions in each of these systems seem to be followed by impairments in the recognition of other people's states (Adolphs, Damasio, Tranel, Cooper, & Damasio, 2000; Calder, Keane, Manes, Antoun, & Young, 2000).

Functions of the motor mirror neuron system

We now start to have a relatively detailed understanding of the properties of mirror neurons in monkeys, and about factors that influence the BOLD activity in the pMNS in humans. In contrast, one of the biggest challenges for research in the next decade will most likely be to reach an understanding of the function of this remarkable system. At present, many speculations exist, most of which are very plausible deductions from properties of the system, but hard evidence supporting the causal relationship between (p)MNS activity and that function remains scarce. What we write

below should thus be seen as a primer for further thoughts and experiments, rather than a definitive guide to the function of this system.

Action recognition

That mirror neurons link the observation/audition of an action to motor programs for the same action suggests that they may play a key role in perceiving what others do and how they are doing it by triggering an internal simulation of the perceived actions (Thioux et al., 2008). Evidence that the MNS indeed contributes to our understanding of other people's actions primarily derives from studies that show that TMS induced perturbations of regions associated with the pMNS or neurological lesions in such regions cause (subtle) deficits in action understanding. Disrupting the ventral premotor cortex using repetitive TMS (rTMS) impairs participants' capacity to judge how heavy a box is when seeing someone else lift the box (Pobric & Hamilton, 2006). Participants that suffer from limb apraxia have difficulties in deciding whether a hand gesture they observe is meaningful or meaningless, with performance in this perceptual task being correlated with their capacity to imitate intransitive gestures (Pazzaglia, Smania, Corato, & Aglioti, 2008b). Lesion analysis showed that patients with limb apraxia who showed more action recognition difficulties were more likely to have lesions in ventral premotor cortex. The fact that many patients with apraxia and lesions in the MNS were still able to perceive some of the gestures correctly shows that the MNS is not the only system that can help recognize actions, but the significant deficits observed in the majority of patients shows that it can significantly contribute to action recognition. In addition, participants with apraxia also have difficulties in recognizing the sound of other people's actions, with those suffering from apraxia of the mouth more impaired in recognizing mouth action sounds, and those suffering from apraxia of the limb more impaired in recognizing hand action sounds (Pazzaglia, Pizzamiglio, Pes, & Aglioti, 2008a), in agreement with the somatotopic organization of the auditory MNS (Gazzola et al., 2006). Another approach to selecting patients with lesions in the frontal pMNS regions is to look for patients with Broca's aphasia. Patients with Broca's aphasia have been shown to be more impaired in sequencing elements of a hand action (e.g. putting four photos taken during a grasping action into the correct order) than performing a similar task on physical phenomena (e.g. ordering photos of a bicycle falling to the ground) (Fazio, Cantagallo, Craighero, D'ausilio, Roy, Pozzo, et al., 2009). Unbiased lesion mapping approaches have also been used to explore the neural basis of action understanding. Kalenine et al. (Kalenine, Buxbaum, & Coslett, 2010) asked 47 patients with left hemisphere strokes and preserved language capacities to determine which of two movies depicted a certain action (e.g. painting a wall). The distractor movie could either be semantically different from the correct movie (e.g. hammering), or could simply differ from the correct movie in terms of kinematics (e.g. moving hand up and down at speed inappropriate for painting). It turned out that patients with lesions overlapping with the parietal pMNS were impaired at telling the difference between the correct and distractor movies when the distractor differed in terms of kinematics. Also lesions along the posterior middle temporal gyrus lead to impairments in the semantic condition, but not in the other one. Finally, the ventral premotor cortex is also involved in mirroring a very specific type of action: facial expressions (van der Gaag et al., 2007), and lesions to this area impair the recognition of facial expressions (Adolphs et al., 2000).

Imitation

Given their propensity to match observed actions with motor programs necessary to execute the same action, mirror neurons have been proposed to play a role in imitation (Brass & Heyes, 2005;

Heyes, 2001; Iacoboni, 2009). fMRI has demonstrated that activity in the ventral premotor cortex (ventral BA44 in particular) during imitation of finger movements is higher than the combination of the activity during the observation and the execution of the same movements (Iacoboni et al., 1999; Molnar-Szakacs, Iacoboni, Koski, & Mazziotta, 2005) and rTMS induced disruption of this region impairs imitation of an action, but not its execution in response to a spatial cue (Heiser, Iacoboni, Maeda, Marcus, & Mazziotta, 2003). Imitation is likely, however, to require more than a simple matching between action perception and action execution (Gergely, Bekkering, & Kiraly, 2002; Tessari, Canessa, Ukmar, & Rumiati, 2007) and brain systems in addition to the MNS are likely to be essential for dynamically controlling the contribution of the MNS to behavior in a social setting (Kokal, Gazzola, & Keysers, 2009).

Ethologists and developmental psychologists have introduced an important distinction between emulation and imitation. Emulation refers to replicating the goal of an observed action without necessarily copying the means. Imitation in the strict sense requires that the detailed movements used to achieve this action be also copied. Considering that the majority of mirror neurons are of the broadly congruent type, observing an action should trigger a varied set of motor programs that would allow the observer to achieve the same goal. Of these, the motor system is likely to favor the most economical for the observer rather than the one chosen by the demonstrator, making emulation a more natural consequence of mirror neurons than imitation. This seems to be the case for both human infants (Bekkering et al., 2000) and monkeys (Subiaul Cantlon, Holloway, & Terrace, 2004), both of which seem to emulate, rather than imitate. True imitation using the MNS would require additional mechanisms that select the motor output of strictly congruent mirror neurons, and the necessity of these additional mechanisms may explain why adult humans can imitate while monkeys very seldom do so—despite the fact that both have mirror neurons.

Language

The MNS has also been considered to play a role in the evolution of language (Rizzolatti & Arbib, 1998). The fact that mirror neurons in the monkey were discovered in F5, which is likely to be the homologue to ventral BA6, BA44 or 45 in humans, all of which are involved in spoken language, supports this idea. Mirror neurons could facilitate language evolution in two ways.

First, the acoustic signal composing speech is ambiguous. A certain sound for instance will be perceived as a/k/ in front of an/a/, but as a/p/ in front of an/u/. The motor theory of speech perception proposes that the human brain resolves this ambiguity, at least in part, by activating the motor programs it would use to produce this sequence of sounds (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). If this motor program involves pushing the tongue on the palate, we will perceive a/k/, if it involves closing the lips, a/p/. In favour of this theory, one recent experiment found for instance that the perception of spoken syllables interferes with the articulation of speech sounds (Yuen, Davis, Brysbaert, & Rastle, 2010). FMRI and TMS studies show that the premotor and motor cortices, respectively, that are involved in producing speech sounds, become reactivated while we listen to the sounds of other people's speech, and rTMS experiments show that interfering with this activity impairs speech perception (Iacoboni, 2008). Importantly, the regions involved correspond to those in which auditory mirror neurons were found in monkeys (Keysers et al., 2003; Kohler et al., 2002), and where human brain activity is increased while hearing and performing goal directed mouth actions (Gazzola et al., 2006) and encompass classical MNS areas.

Secondly, according to embodied semantics, the meaning of action words like "running" might in part be stored in the motor programs we use to run. Indeed, when reading sentences, the premotor cortex becomes activated in a somatotopic fashion, with words like running, grasping and chewing activating regions involved in executing actions with the foot, hand and mouth,

respectively (Aziz-Zadeh et al., 2006; Hauk, Johnsrude, & Pulvermüller, 2004; Pulvermüller, 2005). The fact that these activations fall within regions showing mirror activity to the sight of other people performing similar actions (Aziz-Zadeh et al., 2006) suggests that some neurons might link the sound of action words to the motor representations much like mirror neurons link the sound of the action itself to the motor representations (Gazzola et al., 2006; Keysers et al., 2003; Kohler et al., 2002). In evolution, onomatopoeic words such as “crack” [a nut], which indeed sound like the action, may have served to recruit mirror neurons to convey action meaning through words. Modern language, with its fully arbitrary association of words and meanings may represent an evolution of this system (Keysers, 2011).

Although the MNS could therefore help in language and its evolution, the MNS is obviously not sufficient for language: monkeys have mirror neurons, but no language and many words are not about actions (e.g. ‘blue’) and, therefore, cannot be embodied in our motor cortex. The scope of MNS theories of language are therefore necessarily limited to specific sub-aspects of language (Toni, De Lange, Noordzij, & Hagoort, 2008) and understanding how the MNS interacts with other systems to enable language will remain an important issue.

Prediction

In the Hebbian model presented above, for simplicity’s sake, we considered that while observing oneself while executing an action, activity in the premotor cortex is simultaneous with sensory representations of the same action in the STS. While this is approximately true at the time scale of entire actions, it takes ~200 ms for the causal chain of events linking premotor and STS activity to unfold: premotor activity trigger primary motor activity, which triggers muscle movements, the resulting visual or auditory signals (re-afference) need to go through the many synaptic stages of auditory and visual processing to project onto the synaptic connections in the parietal and premotor cortices. This means that while we reach and grasp a cup in front of us, motor programs for **grasping** will be active while the STS is still sending representations of the **reaching** phase back to parietal and premotor neurons. What gets to wire together is thus not the premotor command for grasping with the vision of grasping, but the command for **grasping** with the vision of the **reaching** that typically precedes grasping by ~200 ms. The Hebbian model thus predicts that mirror activity would anticipate actions about to occur in ~200 ms. This is exactly what seems to be the case (Borroni et al., 2005; Southgate et al., 2009; Urgesi et al., 2010) and has important consequences for the function of the MNS. Instead of retrieving information about the actions that **have** occurred, it seems to trigger motor representations of the actions to come in the next hundreds of milliseconds. These anticipated motor representations then seem to be sent back to the STS (Gazzola and Keysers, 2009; Iacoboni, Koski, Brass, Bekkering, Woods, Dubeau, et al., 2001; Schippers and Keysers, 2011), where they are compared with the future sensory input, generating a prediction error (Gazzola & Keysers, 2009). The MNS might therefore be best understood as a predictive model (Kilner, Friston, & Frith, 2007), that generates hypothesis about people’s future actions and dynamically compares them with people’s next move to generate a dynamically adjusting model of other people’s intentions (i.e. future actions). In line with this interpretation, we would expect that while viewing predictable actions, more information flow should actually travel from premotor to sensory cortices (predictions) than the other way around (prediction errors). For unpredictable actions, the opposite should be true. This is exactly what we found while participants try to guess the meaning of hand gestures. In the unpredictable beginning of a gesture, information flow is stronger from sensory to premotor regions. Later, as gestures become more predictable, information flow predominates in the opposite, premotor to sensory direction (Schippers & Keysers, 2011). In nature, anticipating the behavior of a prey or predator can mean the difference between life and

death. For humans, it can provide a competitive edge in sports (Aglioti, Cesari, Romani, & Urgesi, 2008), but most importantly, it can be key to truly joint actions (Kokal et al., 2009)—for two people to act in synchrony, each partner's brain has to anticipate the actions of the other partner by ~200 ms to have enough time to program and execute his own action in behavioral synchrony with those observed. To act in 'real time' with someone else, our brain actually has to anticipate. Only by doing so can two people lift a set dinner table without tipping over the glasses, skillfully dance together or even just clap in synchrony. Again, information flow from premotor to sensory regions is observed during joint actions (Kokal & Keysers, 2010), and could provide the neural basis for such anticipations. The fact that such sophisticated computations seem to be the outcome of simple Hebbian learning is an elegant property of the synaptic plasticity in the brain.

Learning by observation

Many have argued that the MNS can help the emergence of culture by equipping humans with the capacity to imitate. Incremental culture and technology however have to adopt the **successful** actions of others, but not their unsuccessful attempts. The motor MNS alone cannot perform this filtering, as it would respond equally to successful and unsuccessful actions. The fact that similar systems exist for emotions may resolve this problem. Individual trial-and-error learning occurs when we perform an action and the outcome is more positive than expected, with dopaminergic reward prediction error signals increasing the likelihood of that particular behavior in the future (Schultz, 2006). Conceptually combining MNS for actions and similar systems for emotions means that while observing the actions and outcomes of others, we would vicariously activate motor **and** emotional representation of similar actions and outcomes, respectively. Only those simulated actions that lead to unexpectedly positive simulated outcomes would then trigger vicarious dopaminergic reward prediction signals that would consolidate those simulated behaviors that were successful. Thus, observation learning could actually use the same mechanisms that govern individual learning, but operate on vicarious representations provided by MNS and similar mirror-like mechanisms for vicarious reward (Keysers, 2011).

Support mentalizing

Mentalizing refers to the capacity to **consciously** attribute mental states and beliefs to other individuals and has been associated with activity in the medial prefrontal cortex (Amodio & Frith, 2006). The tasks investigated in that literature differ from those in the literature on the MNS—mentalizing experiments do not typically involve the movies of actions used for MNS experiments, but explicitly encourage participants to *consciously* think about what other people **think**, while MNS experiments involve seeing/hearing the actions of others without being encouraged to think about the thoughts behind the actions (Keysers & Gazzola, 2007). It remains unclear how the MNS contributes to mentalizing. Some propose that the MNS could feed into the distinct mentalizing brain system when the thoughts of others need to be deduced from their actions: much as we can mentalize about our own actions (why does my heart beat stronger each time I see her?), we could mentalize about the vicarious representations of other people's actions, sensations and emotions (Keysers & Gazzola, 2007). Others emphasize that across many studies, the medial prefrontal cortex involved in mentalizing is only seldom found to be activated in the same studies as the MNS, suggesting that these two systems are often rather independent (Van Overwalle & Baetens, 2009). Understanding when and how the MNS and mentalizing brain regions collaborate will be an important question for the future (de Lange, Spronk, Willems, Toni, & Bekkering, 2008; Thioux et al., 2008), and studying the effective connectivity between these brain regions will be essential. In a first step toward that

direction, we studied connectivity across brains while a participant tries to understand the meaning of the gestures of another. We found that regions of both the pMNS and the mentalizing system of the observers carried information about the state of the motor cortices in the gesturer. If the state of these motor cortices is taken to reflect the intentions behind the gestures, that both the mentalizing network and the pMNS of the observer contained information about these states suggests that they resonate with the motor intentions of the sender (Schippers et al., 2010). In the same line, it was demonstrated that empathic accuracy (the ability to accurately judge the emotions and state of mind of another person) is associated with increased activity in both systems (Zaki, Weber, Bolger, & Ochsner, 2009), and mentalizing about another person or about oneself increases the connectivity between the brain regions involved in this ability and the pMNS (Lombardo, M. V., Chakrabarti, Bullmore, Wheelwright, Sadek, Suckling, Consortium, & Baron-Cohen, 2010).

Dysfunctions of the mirror neuron system

Compulsive imitation

Following frontal lobe lesion, some patients demonstrate a tendency to imitate the behaviors of other people, like scratching their forehead, clapping their hands, and so on (De Renzi, Cavalleri, & Facchini, 1996; Lhermitte, 1983; Lhermitte, Pilon, & Serdaru, 1986). The patients persist in imitating the behavior of the experimenter even after being explicitly told to stop doing so. According to one large survey, the phenomenon would be observed in about 40% of patients with frontal lobe lesions, and would virtually never occur as a consequence of post-rolandic brain lesions (De Renzi et al., 1996). Infarct to the anterior cerebral artery resulting in medial frontal lesions seems to be a frequent cause. The fact that most humans do not overtly and compulsively imitate the actions performed by others indicates the existence of a supervisory system in the brain that ensures that of all the premotor programs that mirror neurons activate, only those that are appropriate in a particular situation will be executed while the others are somehow inhibited (Shallice, Burgess, Schon, & Baxter, 1989). The medial frontal lesions inducing compulsive imitation behaviors probably disrupt this system. The idea of inhibitory control finds further support from the fact that fMRI studies that measure an activation of premotor cortices during action observation sometimes simultaneously measure an inhibition of M1, as if to block the motor output of simulation (Gazzola and Keysers, 2009; Gazzola et al., 2007a).

Autism spectrum disorders

A number of studies have been conducted to test whether autism spectrum disorders (ASD) would be characterized by a dysfunction of the MNS. Roughly, the idea is that since the MNS seems to be supporting our ability to understand other individuals' actions and to share their inner experiences, a dysfunction at this level could explain the social difficulties encountered by those with ASD (Iacoboni & Dapretto, 2006; Oberman & Ramachandran, 2007; Rizzolatti, Fabbri-Destro, & Cattaneo, 2009; Williams, Whiten, Suddendorf, & Perrett, 2001).

Several research teams have used EEG to measure Mu-suppression during the observation of hand actions in individuals with ASD and control participants (Box 14.2). Some have reported an absence or a reduction of the Mu wave suppression in ASD, and concluded that the pre-motor cortex was not (or less) active during the observation of hand actions (Bernier, Dawson, Webb, & Murias, 2007; Oberman, Hubbard, Mcleery, Altschuler, Ramachandran, & Pineda, 2005). Other studies, however, found normal mu-suppression in similar conditions (Fan, Decety, Yang, Liu, & Cheng, 2010; Raymaekers Wiersema, & Roeyers, 2009), or when the subject performing the action was familiar to the participant (Oberman, Ramachandran, & Pineda, 2008).

Using fMRI to investigate the MNS activity during the observation of hand actions has not provided further support to the hypothesis of a MNS dysfunction in ASD. In one study, participants with ASD activated the inferior frontal gyrus (BA44) more than controls while observing meaningless gestures (Martineau, Andersson, Barthélémy, Cottier, & Destrieux, 2010). In another study, the authors demonstrated a normal habituation of the BOLD signal when the same action was viewed or executed multiple times (Dinstein, Thomas, Humphreys, Minshew, Behrmann, & Heeger, 2010), suggesting that participants with ASD re-enacted the observed hand actions just like controls. In accordance with this conclusion, at least two studies have also showed behaviorally that the execution of a movement was slowed-down by the concurrent observation of an incongruent movement in autism (Bird, Leighton, Press, & Heyes, 2007b; Spengler, Bird, & Brass, 2010). In one study, the participants with ASD also demonstrated a normal sensitivity to biological motion, with a more pronounced interference effect when observing a human relative to a robotic arm (Bird et al., 2007b).

In general, the research on the observation of hand actions has not lent strong support to the MNS hypothesis of autism. It will be important however to explore the factors that may influence the activation of the MNS in ASD, like for instance the degree of identification with the actor (Oberman et al., 2008). Moreover, some problems have been identified in the timely processing of action plans during the execution of sequences of actions (e.g. reaching, grasping and bringing to the mouth), which might have consequences on the response of the pre-motor cortex during the observation of such complex actions (Cattaneo, Fabbri-Destro, Boria, Pieraccini, Monti, Cossu, et al. 2007). The SMA is an area where an atypical activity has been observed in ASD during the perception of hand actions (Marsh & Hamilton, 2011), but it is unsure whether the difference between groups was the consequence of reduced mirror neuron activity or a difference in the level of activity during the baseline visual condition.

Other studies have investigated the response of the MNS to the observation of facial expressions. In children, a lack of activity in the inferior frontal gyrus pars opercularis was documented and taken as evidence of an MNS dysfunction (Dapretto, Davies, Pfeifer, Scott, Sigman, Bookheimer, et al., 2006). Testing adult participants, we found no evidence for a group difference in this area, nor any other. However, we found that activity in the same portion of the IFG increased with age in ASD. Although the activity was below the normal level for participants in their 20s, older participants activated this area as much as controls when viewing dynamic facial expressions. Moreover, the level of activity in the region was positively correlated with the scores on a social functioning scale (Bastiaansen, Thioux, Nanetti, Van Der Gaag, Ketelaars, Minderaa, et al., 2011). These findings may indicate a delay in the wiring of the mirror neurons involved in the re-enactment of facial movements. Using electromyography to record the spontaneous facial reactions associated with the observation of emotional facial expressions, it was found that the number of facial reactions that were congruent with the observed emotion tended to increase with age in children with ASD, while this number was consistently high in typically developing children of 7 years and older (Beall, Moody, McIntosh, Hepburn, & Reed, 2008). Facial movements have a special status regarding the development of the MNS, since it is not possible to observe one's own facial movements (see "Plasticity and development"). Moreover, faces also have a very special status for children with ASD who tend to look less at faces than typically developing children. Further research into the early development of the MNS for faces may provide valuable insights into the social difficulties experienced by individuals with ASD.

In conclusion, although it is not in its infancy anymore, the research on the possible involvement of the MNS in the aetiology of ASD has struggled to provide consistent evidences. The development of the MNS for faces might be at risk. In general however, it will be important to try

identifying the factors that may influence the MNS response, keeping in mind that these factors might lie outside of the MNS itself, for instance in the attention paid to biological stimuli, and the degree of identification with the actor.

Acknowledgements

The work was supported by grants of the Dutch science foundation (N.W.O.), including NIHC grants to CK (056-13-013, 056-13-017, 452-04-305) and a VENI grant to VG (451-09-006). MT was funded by N.W.O. grant 056-13-017.

References

- Adolphs, R., Damasio, H., Tranel, D., Cooper, G., & Damasio, A. R. (2000). A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *Journal of Neuroscience* 20, 2683–90.
- Aglioti, S. M., Cesari, P., Romani, M., & Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nature Neuroscience* 11, 1109–16.
- Alaerts, K., Heremans, E., Swinnen, S. P. & Wenderoth, N. (2009). How are observed actions mapped to the observer's motor system? Influence of posture and perspective. *Neuropsychologia* 47: 415–22.
- Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience* 7, 268–77.
- Anisfeld, M. 1991. Neonatal imitation. *Developmental Review* 11, 60–97.
- Arnstein, D., Cui, F., Keyzers, C., Maurits, N. M., & Gazzola, V. (2011). μ -suppression during action observation and execution correlates with BOLD in dorsal premotor, inferior parietal, and SI cortices. *Journal of Neuroscience* 31, 14243–9.
- Avenanti, A., Bolognini, N., Maravita, A., & Aglioti, S. M. (2007). Somatic and motor components of action simulation. *Current Biology* 17, 2129–35.
- Avenanti, A., Minio-Paluello, I., Bufalari, I., & Aglioti, S. M. (2009). The pain of a model in the personality of an onlooker: influence of state-reactivity and personality traits on embodied empathy for pain. *NeuroImage* 44, 275–83.
- Aziz-Zadeh, L., Iacoboni, M., Zaidel, E., Wilson, S., & Mazziotta, J. (2004). Left hemisphere motor facilitation in response to manual action sounds. *European Journal of Neuroscience* 19, 2609–12.
- Aziz-Zadeh, L., Sheng, T., Liew, S. L., & Damasio, H. (2011). Understanding otherness: The neural bases of action comprehension and pain empathy in a congenital amputee. *Cerebral Cortex* 22, 811–19.
- Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current Biology* 16, 1818–23.
- Babiloni, C., Babiloni, F., Carducci, F., Cincotti, F., Coccozza, G., Del Percio, C., Moretti, D. V., & Rossini, P. M. (2002). Human cortical electroencephalography (EEG) rhythms during the observation of simple aimless movements: A high-resolution EEG study. *NeuroImage* 17, 559–72.
- Baird, A. D., Scheffer, I. E., & Wilson, S. J. (2011). Mirror neuron system involvement in empathy: a critical look at the evidence. *Social Neuroscience* 6, 327–35.
- Baldissera, F., Cavallari, P., Craighero, L., & Fadiga, L. (2001). Modulation of spinal excitability during observation of hand actions in humans. *European Journal of Neuroscience* 13, 190–4.
- Banissy, M. J., Cohen Kadosh, R., Maus, G. W., Walsh, V., & Ward, J. (2009). Prevalence, characteristics and a neurocognitive model of mirror-touch synaesthesia. *Experimental Brain Research* 198, 261–72.
- Banissy, M. J., Garrido, L., Kusnir, F., Duchaine, B., Walsh, V., & Ward, J. (2011). Superior facial expression, but not identity recognition, in mirror-touch synesthesia. *Journal of Neuroscience* 31, 1820–4.
- Banissy, M. J. & Ward, J. (2007). Mirror-touch synesthesia is linked with empathy. *Nature Neuroscience* 10, 815–6.

- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34, 163–75.
- Bartels, A., Logothetis, N. K., & Moutoussis, K. (2008). fMRI and its interpretations: an illustration on directional selectivity in area V5/MT. *Trends in Neurosciences* 31, 444–53.
- Bastiaansen, J. A., Thioux, M., & Keysers, C. (2009). Evidence for mirror systems in emotions. *Philosophical Transactions of the Royal Society, London, B Biological Science* 364, 2391–404.
- Bastiaansen, J. A., Thioux, M., Nanetti, L., van der Gaag, C., Ketelaars, C., Minderaa, R., & Keysers, C. (2011). Age-related increase in inferior frontal gyrus activity and social functioning in autism spectrum disorder. *Biological Psychiatry* 69, 832–8.
- Beall, P. M., Moody, E. J., McIntosh, D. N., Hepburn, S. L., & Reed, C. L. (2008). Rapid facial reactions to emotional facial expressions in typically developing children and children with autism spectrum disorder. *Journal of Experimental Child Psychology* 101, 206–23.
- Bekkering, H., Wohlschläger, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *Quarterly Journal of Experimental Psychology A* 53, 153–64.
- Bernier, R., Dawson, G., Webb, S., & Murias, M. (2007). EEG mu rhythm and imitation impairments in individuals with autism spectrum disorder. *Brain and Cognition* 64, 228–37.
- Bird, G., Brindley, R., Leighton, J., & Heyes, C. (2007a). General processes, rather than “goals,” explain imitation errors. *Journal of Experimental Psychology: Human Perception and Performance* 33, 1158–69.
- Bird, G., Leighton, J., Press, C., & Heyes, C. (2007b). Intact automatic imitation of human and robot actions in autism spectrum disorders. *Proceedings of the Royal Society: B Biological Sciences* 274, 3027–31.
- Blakemore, S. J., Bristow, D., Bird, G., Frith, C., & Ward, J. (2005). Somatosensory activations during the observation of touch and a case of vision-touch synaesthesia. *Brain* 128, 1571–83.
- Bonini, L., Rozzi, S., Serventi, F. U., Simone, L., Ferrari, P. F., & Fogassi, L. (2010). Ventral premotor and inferior parietal cortices make distinct contribution to action organization and intention understanding. *Cerebral Cortex* 20, 1372–85.
- Bonini, L., Serventi, F. U., Simone, L., Rozzi, S., Ferrari, P. F., & Fogassi, L. (2011). Grasping neurons of monkey parietal and premotor cortices encode action goals at distinct levels of abstraction during complex action sequences. *Journal of Neuroscience* 31, 5876–86.
- Borroni, P., Montagna, M., Cerri, G., & Baldissera, F. (2005). Cyclic time course of motor excitability modulation during the observation of a cyclic hand movement. *Brain Research* 1065, 115–24.
- Brass, M., Bekkering, H., Wohlschläger, A., & Prinz, W. (2000). Compatibility between observed and executed finger movements: comparing symbolic, spatial, and imitative cues. *Brain and Cognition* 44, 124–43.
- Brass, M., & Heyes, C. (2005). Imitation: is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Science* 9, 489–95.
- Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R. J., Zilles, K., Rizzolatti, G., & Freund, H. J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience* 13, 400–4.
- Caetano, G., Jousmaki, V., & Hari, R. (2007). Actor’s and observer’s primary motor cortices stabilize similarly after seen or heard motor actions. *Proceedings of the National Academy of Science, USA* 104, 9058–62.
- Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience* 3, 1077–8.
- Calmels, C., Holmes, P., Jarry, G., Hars, M., Lopez, E., Paillard, A., & Stam, C. J. (2006). Variability of EEG synchronization prior to and during observation and execution of a sequential finger movement. *Human Brain Mapping* 27, 251–66.

- Calvo-Merino, B., Glaser, D. E., Grezes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: an fMRI study with expert dancers. *Cerebral Cortex* 15, 1243–9.
- Caspers, S., Zilles, K., Laird, A. R., & Eickhoff, S. B. (2010). ALE meta-analysis of action observation and imitation in the human brain. *NeuroImage* 50, 1148–67.
- Catmur, C., Walsh, V., & Heyes, C. (2007). Sensorimotor learning configures the human mirror system. *Current Biology* 17, 1527–31.
- Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., & Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of the National Academy of Science, USA* 104, 17825–30.
- Chakrabarti, B., Bullmore, E. & Baron-Cohen, S. (2006). Empathizing with basic emotions: common and discrete neural substrates. *Social Neuroscience* 1, 364–84.
- Chong, T. T., Cunnington, R., Williams, M. A., Kanwisher, N., & Mattingley, J. B. (2008). fMRI adaptation reveals mirror neurons in human inferior parietal cortex. *Current Biology* 18, 1576–80.
- Clark, S., Tremblay, F., & Ste-Marie, D. (2004). Differential modulation of corticospinal excitability during observation, mental imagery and imitation of hand actions. *Neuropsychologia* 42, 105–12.
- Cochin, S., Barthelemy, C., Lejeune, B., Roux, S., & Martineau, J. (1998). Perception of motion and qEEG activity in human adults. *Electroencephalography and Clinical Neurophysiology* 107, 287–95.
- Cochin, S., Barthelemy, C., Roux, S., & Martineau, J. (1999). Observation and execution of movement: similarities demonstrated by quantified electroencephalography. *European Journal of Neuroscience* 11, 1839–42.
- Cross, E. S., Hamilton, A. F., & Grafton, S. T. (2006). Building a motor simulation de novo: observation of dance by dancers. *NeuroImage* 31, 1257–67.
- Danziger, N., Faillenot, I., & Peyron, R. (2009). Can we share a pain we never felt? Neural correlates of empathy in patients with congenital insensitivity to pain. *Neuron* 61, 203–12.
- Danziger, N., Prkachin, K. M., & Willer, J.-C. (2006). Is pain the price of empathy? The perception of others' pain in patients with congenital insensitivity to pain. *Brain* 129, 2494–507.
- Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., & Iacoboni, M. (2006). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* 9, 28–30.
- Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology* 10, 85–104.
- de Lange, F. P., Spronk, M., Willems, R. M., Toni, I., & Bekkering, H. (2008). Complementary systems for understanding action intentions. *Current Biology* 18, 454–7.
- de Renzi, E., Cavalleri, F., & Facchini, S. (1996). Imitation and utilization behavior. *Journal of Neurology, Neurosurgery and Psychiatry* 61, 396–400.
- del Giudice, M., Manera, V., & Keysers, C. (2009). Programmed to learn? The ontogeny of mirror neurons. *Developmental Science* 12, 350–63.
- Dinstein, I., Hasson, U., Rubin, N., & Heeger, D. J. (2007). Brain areas selective for both observed and executed movements. *Journal of Neurophysiology* 98, 1415–27.
- Dinstein, I., Thomas, C., Humphreys, K., Minshew, N., Behrmann, M., & Heeger, D. J. (2010). Normal movement selectivity in autism. *Neuron* 66, 461–9.
- Ebisch, S. J., Perrucci, M. G., Ferretti, A., del Gratta, C., Romani, G. L., & Gallese, V. (2008). The sense of touch: Embodied simulation in a visuotactile mirroring mechanism for observed animate or inanimate touch. *Journal of Cognitive Neuroscience* 20, 1611–23.
- Eisenberger, N. I. (2012). The pain of social disconnection: examining the shared neural underpinnings of physical and social pain. *Nature Reviews Neuroscience* 13, 421–34.
- Etzel, J. A., Gazzola, V., & Keysers, C. (2008). Testing simulation theory with cross-modal multivariate classification of fMRI data. *PLoS One* 3, e3690.

- Evangelidou, M. N., Raos, V., Galletti, C., & Savaki, H. E. (2009). Functional imaging of the parietal cortex during action execution and observation. *Cerebral Cortex* 19, 624–39.
- Fadiga, L., Craighero, L., & Olivier, E. (2005). Human motor cortex excitability during the perception of others' action. *Current Opinion in Neurobiology* 15, 213–18.
- Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology* 73, 2608–11.
- Fan, Y.-T., Decety, J., Yang, C.-Y., Liu, J.-L., & Cheng, Y. (2010). Unbroken mirror neurons in autism spectrum disorders. *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 51, 981–8.
- Fazio, P., Cantagallo, A., Craighero, L., D'Ausilio, A., Roy, A. C., Pozzo, T., Calzolari, F., Granieri, E., & Fadiga, L. (2009). Encoding of human action in Broca's area. *Brain* 132, 1980–8.
- Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience* 17, 1703–14.
- Filimon, F., Nelson, J. D., Hagler, D. J., & Sereno, M. I. (2007). Human cortical representations for reaching: mirror neurons for execution, observation, and imagery. *NeuroImage* 37, 1315–28.
- Fujii, N., Hihara, S., & Iriki, A. (2008). Social cognition in premotor and parietal cortex. *Social Neuroscience* 3, 250–60.
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119(Pt 2), 593–609.
- Gastaut, H. J., & Bert, J. (1954). EEG changes during cinematographic presentation; moving picture activation of the EEG. *Electroencephalography and Clinical Neurophysiology* 6, 433–44.
- Gazzola, V., Aziz-Zadeh, L., & Keysers, C. (2006). Empathy and the somatotopic auditory mirror system in human. *Current Biology* 16, 1824–9.
- Gazzola, V., & Keysers, C. (2009). The observation and execution of actions share motor and somatosensory voxels in all tested subjects: single-subject analyses of unsmoothed fMRI data. *Cerebral Cortex* 19, 1239–55.
- Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007a). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage* 35, 1674–84.
- Gazzola, V., van der Worp, H., Mulder, T., Wicker, B., Rizzolatti, G., & Keysers, C. (2007b). Aphasics born without hands mirror the goal of hand actions with their feet. *Current Biology* 17, 1235–40.
- Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature* 415, 755.
- Graziano, M. S., Taylor, C. S., & Moore, T. (2002). Complex movements evoked by microstimulation of precentral cortex. *Neuron* 34, 841–51.
- Grezes, J., Armony, J. L., Rowe, J., & Passingham, R. E. (2003). Activations related to “mirror” and “canonical” neurones in the human brain: an fMRI study. *NeuroImage* 18, 928–37.
- Hamilton, A. F., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. *Journal of Neuroscience* 26, 1133–7.
- Hamilton, A. F., & Grafton, S. T. (2008). Action outcomes are represented in human inferior frontoparietal cortex. *Cerebral Cortex* 18, 1160–8.
- Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., & Rizzolatti, G. (1998). Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proceedings of the National Academy of Sciences, USA* 95, 15061–5.
- Haslinger, B., Erhard, P., Altenmüller, E., Schroeder, U., Boecker, H., & Ceballos-Baumann, A. O. (2005). Transmodal sensorimotor networks during action observation in professional pianists. *Journal of Cognitive Neuroscience* 17, 282–93.
- Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron* 41, 301–7.
- Heiser, M., Iacoboni, M., Maeda, F., Marcus, J., & Mazziotta, J. C. (2003). The essential role of Broca's area in imitation. *European Journal of Neuroscience* 17, 1123–8.

- Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Science* 5, 253–61.
- Hutchison, W. D., Davis, K.D., Lozano, A. M., Tasker, R. R., & Dostrovsky, J. O. (1999). Pain-related neurons in the human cingulate cortex. *Nature Neuroscience* 2, 403–5.
- Iacoboni, M. (2008). The role of premotor cortex in speech perception: evidence from fMRI and rTMS. *Journal of Physiology, Paris* 102, 31–4.
- Iacoboni, M. (2009). Neurobiology of imitation. *Current Opinion in Neurobiology* 19, 661–5.
- Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience* 7, 942–51.
- Iacoboni, M., Koski, L. M., Brass, M., Bekkering, H., Woods, R. P., Dubeau, M. C., Mazziotta, J. C., & Rizzolatti, G. (2001). Reafferent copies of imitated actions in the right superior temporal cortex. *Proceedings of the National Academy of Science, USA* 98, 13995–9.
- Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology* 3, e79.
- Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science* 286, 2526–8.
- Ishida, H., Nakajima, K., Inase, M., & Murata, A. (2010). Shared mapping of own and others' bodies in visuotactile bimodal area of monkey parietal cortex. *Journal of Cognitive Neuroscience* 22, 83–96.
- Jabbi, M., & Keysers, C. (2008). Inferior frontal gyrus activity triggers anterior insula response to emotional facial expressions. *Emotion* 8, 775–80.
- Jabbi, M., Swart, M., & Keysers, C. (2007). Empathy for positive and negative emotions in the gustatory cortex. *NeuroImage* 34, 1744–53.
- Kalenine, S., Buxbaum, L. J., & Coslett, H. B. (2010). Critical brain regions for action recognition: lesion symptom mapping in left hemisphere stroke. *Brain* 133, 3269–80.
- Kanakogi, Y., & Itakura, S. (2011). Developmental correspondence between action prediction and motor ability in early infancy. *Nature Communications* 2, 341.
- Keysers, C. (2011). *The Empathic Brain*. Amsterdam: Social Brain Press.
- Keysers, C., & Gazzola, V. (2007). Integrating simulation and theory of mind: from self to social cognition. *Trends in Cognitive Science* 11, 194–6.
- Keysers, C., & Gazzola, V. (2009). Expanding the mirror: vicarious activity for actions, emotions, and sensations. *Current Opinion in Neurobiology* 19, 666–71.
- Keysers, C., & Gazzola, V. (2010). Social neuroscience: mirror neurons recorded in humans. *Current Biology* 20, R353–4.
- Keysers, C., Kaas, J. H., & Gazzola, V. (2010). Somatosensation in social perception. *Nature Reviews Neuroscience* 11, 417–28.
- Keysers, C., Kohler, E., Umiltà, M. A., Nanetti, L., Fogassi, L., & Gallese, V. (2003). Audiovisual mirror neurons and action recognition. *Experimental Brain Research* 153, 628–36.
- Keysers, C., & Perrett, D. I. (2004). Demystifying social cognition: A Hebbian perspective. *Trends in Cognitive Science* 8, 501–7.
- Keysers, C., Wicker, B., Gazzola, V., Anton, J. L., Fogassi, L., & Gallese, V. (2004). A touching sight: SII/PV activation during the observation and experience of touch. *Neuron* 42, 335–46.
- Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Process* 8, 159–66.
- Kilner, J. M., Neal, A., Weiskopf, N., Friston, K. J., & Frith, C. D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. *Journal of Neuroscience* 29, 10153–9.
- Kilner, J. M., Paulignan, Y., & Blakemore, S. J. (2003). An interference effect of observed biological movement on action. *Current Biology* 13, 522–5.
- Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297, 846–8.

- Kokal, I., Gazzola, V., & Keysers, C. (2009). Acting together in and beyond the mirror neuron system. *NeuroImage* 47, 2046–56.
- Kokal, I., & Keysers, C. (2010). Granger causality mapping during joint actions reveals evidence for forward models that could overcome sensory-motor delays. *PLoS One* 5, e13507.
- Lahav, A., Saltzman, E., & Schlaug, G. (2007). Action representation of sound: audiomotor recognition network while listening to newly acquired actions. *Journal of Neuroscience* 27, 308–14.
- Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage* 54, 2492–502.
- Lepage, J. F., & Theoret, H. (2006). EEG evidence for the presence of an action observation-execution matching system in children. *European Journal of Neuroscience* 23, 2505–10.
- Lhermitte, F. (1983). 'Utilization behavior' and its relation to lesions of the frontal lobes. *Brain* 106(Pt 2), 237–55.
- Lhermitte, F., Pillon, B., & Serdaru, M. (1986). Human autonomy and the frontal lobes. Part I: Imitation and utilization behavior: a neuropsychological study of 75 patients. *Annals of Neurology* 19, 326–34.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review* 74, 431–61.
- Lingnau, A., Gesierich, B., & Caramazza, A. (2009). Asymmetric fMRI adaptation reveals no evidence for mirror neurons in humans. *Proceedings of the National Academy of Science, USA* 106, 9925–30.
- Lippert, M. T., Steudel, T., Ohl, F., Logothetis, N. K., & Kayser, C. (2010). Coupling of neural activity and fMRI-BOLD in the motion area MT. *Magnetic Resonance Imaging* 28, 1087–94.
- Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., Consortium, M. A., & Baron-Cohen, S. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience* 22, 1623–35.
- Lui, F., Buccino, G., Duzzi, D., Benuzzi, F., Crisi, G., Baraldi, P., Nichelli, P., Porro, C. A. & Rizzolatti, G. (2008). Neural substrates for observing and imagining non-object-directed actions. *Social Neuroscience* 3, 261–75.
- Marsh, L. E., & Hamilton, A. F. D. C. (2011). Dissociation of mirroring and mentalizing systems in autism. *NeuroImage* 56, 1511–19.
- Martineau, J., Andersson, F., Barthélémy, C., Cottier, J-P., & Destrieux, C. (2010). Atypical activation of the mirror neuron system during perception of hand motion in autism. *Brain Research* 1320C, 168–75.
- Meltzoff, A. N., & Decety, J. (2003). What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society, London, B Biological Sciences* 358, 491–500.
- Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science* 198, 75–8.
- Molnar-Szakacs, I., Iacoboni, M., Koski, L., & Mazziotta, J. C. (2005). Functional segregation within pars opercularis of the inferior frontal gyrus: evidence from fMRI studies of imitation and action observation. *Cerebral Cortex* 15, 986–94.
- Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology* 20, 750–6.
- Muthukumaraswamy, S. D., & Johnson, B. W. (2004a). Changes in rolandic mu rhythm during observation of a precision grip. *Psychophysiology* 41, 152–6.
- Muthukumaraswamy, S. D., & Johnson, B. W. (2004b). Primary motor cortex activation during action observation revealed by wavelet analysis of the EEG. *Clinical Neurophysiology* 115, 1760–6.
- Muthukumaraswamy, S. D., Johnson, B. W., & McNair, N. A. (2004). Mu rhythm modulation during observation of an object-directed grasp. *Brain Research, Cognitive Brain Research* 19, 195–201.
- Nelissen, K., Borra, E., Gerbella, M., Rozzi, S., Luppino, G., Vanduffel, W., Rizzolatti, G., & Orban, G. A. (2011). Action observation circuits in the macaque monkey cortex. *Journal of Neuroscience* 31, 3743–56.

- Oberman, L. M., Hubbard, E. M., McCleery, J. P., Altschuler, E. L., Ramachandran, V. S., & Pineda, J. A. (2005). EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Brain Research, Cognitive Brain Research* 24, 190–8.
- Oberman, L. M., & Ramachandran, V. S. (2007). The simulating social mind: the role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. *Psychological Bulletin* 133, 310–27.
- Oberman, L. M., Ramachandran, V. S., & Pineda, J. A. (2008). Modulation of mu suppression in children with autism spectrum disorders in response to familiar or unfamiliar stimuli: the mirror neuron hypothesis. *Neuropsychologia* 46, 1558–65.
- Oosterhof, N. N., Wiggett, A. J., Diedrichsen, J., Tipper, S. P., & Downing, P. E. (2010). Surface-based information mapping reveals crossmodal vision-action representations in human parietal and occipitotemporal cortex. *Journal of Neurophysiology* 104, 1077–89.
- Oztop, E., & Arbib, M. A. (2002). Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics* 87, 116–40.
- Pazzaglia, M., Pizzamiglio, L., Pes, E., & Aglioti, S. M. (2008a). The sound of actions in apraxia. *Current Biology* 18, 1766–72.
- Pazzaglia, M., Smania, N., Corato, E., & Aglioti, S. M. (2008b). Neural underpinnings of gesture discrimination in patients with limb apraxia. *Journal of Neuroscience* 28, 3030–41.
- Pfeifer, J. H., Iacoboni, M., Mazziotta, J. C., & Dapretto, M. (2008). Mirroring others' emotions relates to empathy and interpersonal competence in children. *NeuroImage* 39, 2076–85.
- Pineda, J. A. (2005). The functional significance of mu rhythms: translating “seeing” and “hearing” into “doing”. *Brain Research, Brain Research Review* 50, 57–68.
- Pizzamiglio, L., Aprile, T., Spitoni, G., Pitzalis, S., Bates, E., D'Amico, S., & di Russo, F. (2005). Separate neural systems for processing action- or non-action-related sounds. *NeuroImage* 24, 852–61.
- Pobric, G., & Hamilton, A. F. (2006). Action understanding requires the left inferior frontal cortex. *Current Biology* 16, 524–9.
- Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience* 6, 576–82.
- Raos, V., Evangelidou, M. N., & Savaki, H. E. (2004). Observation of action: grasping with the mind's hand. *NeuroImage* 23, 193–201.
- Raymaekers, R., Wiersema, J. R., & Roeyers, H. (2009). EEG study of the mirror neuron system in children with high functioning autism. *Brain Research* 1304, 113–21.
- Ricciardi, E., Bonino, D., Sani, L., Vecchi, T., Guazzelli, M., Haxby, J. V., Fadiga, L., & Pietrini, P. (2009). Do we really need vision? How blind people “see” the actions of others. *Journal of Neuroscience* 29, 9719–24.
- Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neuroscience* 21, 188–94.
- Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey. II. Area F5 and the control of distal movements. *Experimental Brain Research* 71, 491–507.
- Rizzolatti, G., Fabbri-Destro, M., & Cattaneo, L. (2009). Mirror neurons and their clinical relevance. *Nature Clinical Practice Neurology* 5, 24–34.
- Rossi, S., Tecchio, F., Pasqualetti, P., Olivelli, M., Pizzella, V., Romani, G. L., Passero, S., Battistini, N., & Rossini, P. M. (2002). Somatosensory processing during movement observation in humans. *Clinical Neurophysiology* 113, 16–24.
- Rozzi, S., Calzavara, R., Belmalih, A., Borra, E., Gregoriou, G. G., Matelli, M., & Luppino, G. (2006). Cortical connections of the inferior parietal cortical convexity of the macaque monkey. *Cerebral Cortex* 16, 1389–417.
- Rozzi, S., Ferrari, P. F., Bonini, L., Rizzolatti, G., & Fogassi, L. (2008). Functional organization of inferior parietal lobule convexity in the macaque monkey: electrophysiological characterization of motor,

- sensory and mirror responses and their correlation with cytoarchitectonic areas. *European Journal of Neuroscience* 28, 1569–88.
- Schippers, M. B., Gazzola, V., Goebel, R., & Keysers, C. (2009). Playing charades in the fMRI: are mirror and/or mentalizing areas involved in gestural communication? *PLoS One* 4, e6801.
- Schippers, M. B., & Keysers, C. (2011). Mapping the flow of information within the putative mirror neuron system during gesture observation. *NeuroImage* 57, 37–44.
- Schippers, M. B., Roebroek, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flow from one brain to another during gestural communication. *Proceedings of the National Academy of Sciences, USA* 107, 9388–93.
- Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology* 57, 87–115.
- Shallice, T., Burgess, P. W., Schon, F., & Baxter, D. M. (1989). The origins of utilization behavior. *Brain* 112(Pt 6), 1587–98.
- Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus vs. ventromedial prefrontal lesions. *Brain* 132, 617–27.
- Shimada, S., & Hiraki, K. (2006). Infant's brain responses to live and televised action. *NeuroImage* 32, 930–9.
- 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, 1157–62.
- Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition* 96, B1–11.
- Southgate, V., Johnson, M. H., Osborne, T., & Csibra, G. (2009). Predictive motor activation during action observation in human infants. *Biological Letters* 5, 769–72.
- Spengler, S., Bird, G., & Brass, M. (2010). Hyperimitation of actions is related to reduced understanding of others' minds in autism spectrum conditions. *Biological Psychiatry* 68, 1148–55.
- Spunt, R. P., Satpute, A. B., & Lieberman, M. D. (2011). Identifying the what, why, and how of an observed action: an fMRI study of mentalizing and mechanizing during action observation. *Journal of Cognitive Neuroscience* 23, 63–74.
- Subiaul, F., Cantlon, J. F., Holloway, R. L., & Terrace, H. S. (2004). Cognitive imitation in rhesus macaques. *Science* 305, 407–10.
- Tessari, A., Canessa, N., Ukmar, M., & Rumiati, R. I. (2007). Neuropsychological evidence for a strategic control of multiple routes in imitation. *Brain* 130, 1111–26.
- Thioux, M., Gazzola, V., & Keysers, C. (2008). Action understanding: how, what and why. *Current Biology* 18, R431–4.
- Toni, I., de Lange, F. P., Noordzij, M. L., & Hagoort, P. (2008). Language beyond action. *Journal of Physiology, Paris* 102, 71–9.
- Turella, L., Erb, M., Grodd, W., & castiello, U. (2009). Visual features of an observed agent do not modulate human brain activity during action observation. *NeuroImage* 46, 844–53.
- Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I know what you are doing. a neurophysiological study. *Neuron* 31, 155–65.
- Urgesi, C., Candidi, M., Fabbro, F., Romani, M., & Aglioti, S. M. (2006). Motor facilitation during action observation: topographic mapping of the target muscle and influence of the onlooker's posture. *European Journal of Neuroscience* 23, 2522–30.
- Urgesi, C., Maieron, M., Avenanti, A., Tidoni, E., Fabbro, F., & Aglioti, S. M. (2010). Simulating the future of actions in the human corticospinal system. *Cerebral Cortex* 20, 2511–21.
- van der Gaag, C., Minderaa, R. B., & Keysers, C. (2007). Facial expressions: what the mirror neuron system can and cannot tell us. *Social Neuroscience* 2, 179–222.

- van Elk, M., van Schie, H. T., Hunnius, S., Vesper, C., & Bekkering, H. (2008). You'll never crawl alone: neurophysiological evidence for experience-dependent motor resonance in infancy. *NeuroImage* **43**, 808–14.
- van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage* **48**, 564–84.
- Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron* **40**, 655–64.
- Williams, J. H., Whiten, A., Suddendorf, T., & Perrett, D. I. (2001). Imitation, mirror neurons and autism. *Neuroscience and Biobehavioral Reviews* **25**, 287–95.
- Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition* **69**, 1–34.
- Yuen, I., Davis, M. H., Brysbaert, M., & Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Sciences, USA* **107**, 592–7.
- Zaki, J., Weber, J., Bolger, N., & Ochsner, K. (2009). The neural bases of empathic accuracy. *Proceedings of the National Academy of Sciences, USA* **106**, 11382–7.

The mirror mechanism: Understanding others from the inside

Giacomo Rizzolatti and Maddalena Fabbri-Destro

Introduction

A fundamental issue of cognitive neuroscience is to understand how we are able to comprehend actions and intentions of others. A striking characteristic of this process is that typically we interpret others' behavior as a mark of something rather insubstantial as mental activity. The behaviors of others are clues to their goals and intention waiting to be recognized.

The most common interpretation of this ability is that it derives from the observers' capacity to logically infer others' internal mental states from the available sensory information, and to ascribe to them a causal role in generating the observed behavior (e.g. Carruthers & Smith, 1996; Malle, Moses, & Baldwin, 2001). An alternative view is that advanced by phenomenologists (see Merleau-Ponty, 1962). According to them we make sense, in normal everyday life, of others' behavior as if it was our own. We make sense of others without resorting to inferential processes, but relying instead on an immediate and direct understanding of *what* others do and *why* they are doing.

It would be dogmatic to assume that there is one and only one way in which we recognize actions and intentions of others. There are undoubtedly various ways in which we can solve this problem. A point, however, must be stressed. There is, a fundamental difference from the direct way of others' understanding, proposed by the phenomenologists, and theories relying on inferential reasoning. Inferential reasoning does not differentiate, a priori, the understanding of physical phenomena from that of actions done by of our conspecifics. The inferential understanding of an apple falling from a tree does not include empathy with the apple. The same is true if you apply the inferential mechanism to the understanding of human behavior. You perfectly understand what is going on, but not a fundamental aspect of it: the feeling of what occurred to another individual. No empathy between you and the other person is present. In contrast, this empathic aspect is captured by the phenomenological explanation of others' behavior. Not only what an individual does is understood, but also, within limits, what he feels in doing it.

Neurobiological evidence for a direct understanding of behavior of others has been provided by the discovery of a specific class of neurons in the monkey premotor cortex. These neurons—called mirror neurons—discharge both when people perform a given motor act and when they observe someone else performing a similar motor act (di Pellegrino, Fadiga, Fogassi L., Gallese, & Rizzolatti, 1992; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti et al., 1996a). Subsequently, mirror neurons have been also found in the parietal cortex of the monkey (Fogassi, Ferrari, Gesierich, Rozzi, Chersi, & Rizzolatti, 2005), in motor and visceromotor human brain centers (see Fabbri-Destro & Rizzolatti, 2008; Rizzolatti & Craighero, 2004) and in song-producing motor areas of birds (Keller & Hahnloser, 2009; Prather, Peters, Nowicki, & Mooney, 2008).

Mirror neurons have been demonstrated by single neuron studies (monkeys, birds, and some human data), as well as by non-invasive techniques—EEG, MEG, PET, fMRI, TMS (mostly human

data). All of them have in common a basic mechanism (the “mirror mechanism”) that transforms sensory representations of actions into a motor format. According to its anatomical location, this mechanism subserves different functions, ranging from the recognition of song of conspecifics in birds to action understanding and empathy in humans.

We shall describe first the functional proprieties of monkey mirror neurons. We shall then discuss the mirror mechanism in humans by presenting data that suggest that the motor encoding of actions is critical for perceiving the motor acts of others not only in terms of their goals, but also in terms of their intention. We shall conclude with some considerations on the relations between the mirror mechanism and autism.

The mirror mechanism in the monkey

Mirror neurons are **motor neurons**. Therefore, to understand their functions one must first have clear the basic proprieties of the motor areas where the mirror neurons are located. Before describing the properties of mirror neurons we will briefly summarize the properties of frontal and parietal motor areas

The organization of the motor cortex in the monkey

The agranular (motor) frontal cortex of the macaque monkey occupies the caudal part of the frontal lobe. According to the classical subdivision of Brodmann (1909), this cortex is formed by two large cytoarchitectonic areas—4 and 6. In recent years, the combination of classic architectonic and neurochemical techniques has proven to be extremely useful for a more objective assessment of architectonic areas (Belmalih, Borra, Contini, Gerbella, Rozzi, & Luppino, 2007; Geyer, Matelli, Luppino, & Zilles, 2000).

This combination, applied to the motor cortex, resulted in the parcellation presented in Figure 15. 1 (Belmalih, Borra, Contini, Gerbella, Rozzi, & Luppino, 2009; Matelli, Luppino, & Rizzolatti, 1985, 1991). In this parcellation, area F1 roughly corresponds to area 4 (primary motor cortex), whereas each of the three main sectors of Brodmann area 6 (mesial, dorsal, and ventral area 6) are formed by a caudal and a rostral subdivision. Specifically, the mesial sector is formed by areas F3 (SMA) and F6 (pre SMA), the dorsal sector by F2 and F7 (together called dorsal premotor cortex) and F4 and F5 (together called ventral premotor cortex).

The motor areas located rostrally to the primary motor area F1 can be grouped into two major classes (see Rizzolatti & Luppino 2001). The caudal premotor areas F2, F3, F4, F5 constitute the first class. They receive their main input from the parietal lobe. The rostral premotor areas F6, and F7 constitute the second class. They receive their main input from the prefrontal cortex.

Studies on the organization of the parieto-frontal connections showed that motor and parietal areas are connected one with another in a very specific way (Rizzolatti, Luppino, & Matelli, 1998). In particular areas of the superior parietal lobule (SPL) are connected with dorsal and mesial premotor areas, while areas of the inferior parietal lobule (IPL) are connected with ventral premotor areas. Prefrontal projections to the motor cortex are primarily directed to rostral premotor areas (Gerbella, Belmalih, Borra, Rozzi, & Luppino, 2010; Lu, Preston, & Strick, 1994; Luppino, Matelli, Camarda, & Rizzolatti, 1993; Rizzolatti & Luppino 2001). Thus, caudal and rostral premotor areas play a different functional role in motor control. Caudal premotor areas, together with parietal areas, are involved in transformation of sensory information into potential motor acts. In contrast, rostral premotor areas are fundamental for determining, according to the external and internal contingencies, when potential motor acts will be executed.

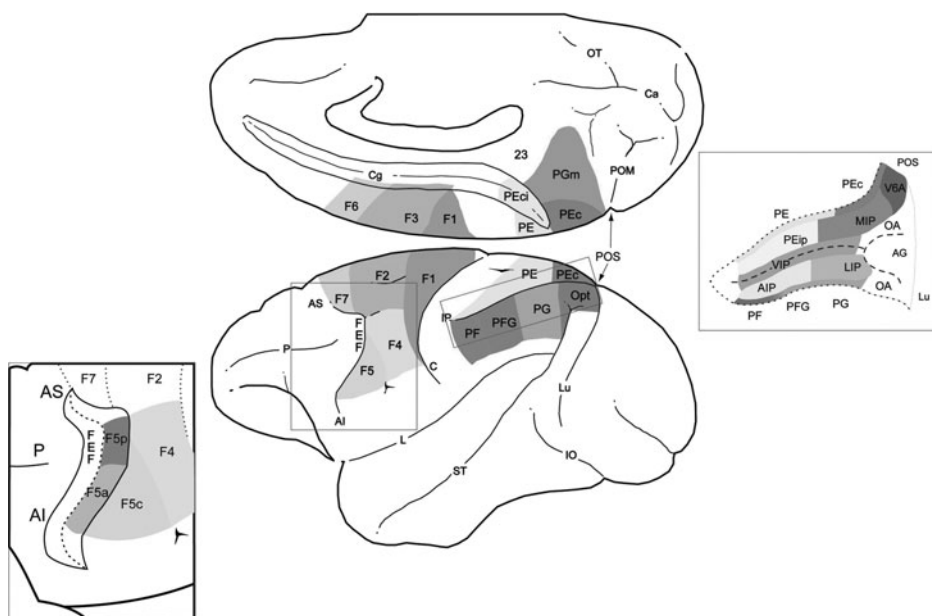


Figure 15.1 Mesial and lateral views of the macaque brain. The Figure shows the cytoarchitectonic parcellation of the frontal motor cortex (areas indicated with F and Arabic numbers) and of the parietal lobe (areas indicated with P and Latin letters). Areas buried within the intraparietal sulcus are shown on the right in an unfolded view of the sulcus. The left inset shows the subareas forming area F5. AIP, anterior intraparietal area; As, superior arcuate sulcus; Ai inferior arcuate sulcus; C, central sulcus; Ca, calcarine fissure; CG, cingulate cortex; FEF, frontal eye field; IP, intraparietal sulcus; L, lateral sulcus; LIP, lateral intraparietal area; MIP, medial intraparietal area; Lu, lunate sulcus; P, principal sulcus; POs, parieto-occipital sulcus; STS, superior temporal sulcus. See also Plate 6.

Premotor area F5

Area F5 forms the rostral part of the ventral premotor cortex. There is evidence that F5, originally considered as a single entity, is constituted by three sectors (Belmalih et al. 2009). One called F5 “convexity” (F5c), is located on the post-arcuate cortex convexity. The others, F5 “posterior” and “anterior” (F5p and F5a, respectively), lie within the post-arcuate bank, at different antero-posterior levels (Figure 15. 1).

Single neuron recording studies from area F5 revealed that most of its neurons code specific goals, rather than individual movements (Rizzolatti, Camarda, Fogassi, Gentilucci, Luppino, & Matelli, 1988; Umiltà, Escola, Intskirveli, Grammont, Rochat, Caruana, et al., 2008). Using the effective motor act as the classification criterion, F5 neurons were subdivided into various classes. Among them, the most represented are: “grasping,” “holding,” “tearing,” and “manipulating” neurons. Neurons of a given class respond weakly or not at all when similar movements are executed for a different goal.

F5 contains neurons responding not only during execution of motor acts, but also to the presentation of visual stimuli (Rizzolatti et al., 1988). F5 visuo-motor neurons fall into two classes—"canonical" neurons and "mirror" neurons.

Canonical neurons, mostly located in area F5p, discharge to the presentation of 3D objects (Murata, Fadiga, Fogassi, Gallese, Raos, & Rizzolatti, 1997; Raos, Umiltà, Murata, Fogassi, & Gallese, 2006). They have been systematically studied using a paradigm that allows one to separate activity related to object presentation, action preparation, and action execution. The most interesting result was that the majority of canonical neurons responded selectively to objects of a certain size, shape, and orientation. Typically, visual specificity was congruent with motor specificity (Murata et al., 1997).

The second category of F5 visuomotor neurons corresponds to mirror neurons. They are mostly located in F5c (Figure 15.1). This area contains a hand/mouth motor representation, with the hand represented mostly in its medial part, and the mouth in the lateral one.

Parietal areas PFG, AIP, LIP, and the superior temporal sulcus region

It is well known that the anterior region of the superior temporal sulcus (STSa) contains neurons responding to the observation of arm and hand movements (Barraclough, Xiao, Oram, & Perrett, 2006; Perrett et al., 1989). STS neurons do not discharge, however, in association with motor activity. They cannot be, therefore considered, mirror neurons.

Recently, an fMRI study was carried out to define the pathways that link STS with F5. This study revealed that during action observation three regions were activated: STS, IPL, and the cortex around the arcuate sulcus. A subsequent ROIs analysis showed that various sectors of STS (both in its upper and lower banks) and two areas of IPL became active (see Figure 15.2; Nelissen et al., 2011;). In agreement with these findings, single neuron recording demonstrated the presence of mirror neurons in areas PFG (Fogassi et al., 2005; Rozzi, Ferrari, Bonini, Rizzolatti, & Fogassi, 2008) and AIP (S. Rozzi, unpublished data).

Another area of the parietal lobe also appears to host mirror neurons—area lateral intraparietal area (LIP), which forms with the frontal eye field a circuit involved in the organization of eye movements. The properties of mirror neurons located in area LIP have been described by Shepherd and colleagues (Shepherd, Klein, Deaner, & Platt, 2009). They found that a set of LIP neurons fired both when the monkey looked in the neuron-preferred direction and when it saw another monkey looking in the same direction. Interestingly, another subset of LIP neurons that discharged when the recorded monkey looked toward a certain direction was, in contrast, suppressed when the observed monkey looked in the same direction. The authors suggested that LIP mirror neurons contribute to the sharing of observed attention and might play role in imitative behavior.

Functional proprieties of mirror neurons: understanding the goal of others' motor acts

Mirror neurons discharge both when the monkey performs a hand goal-directed motor act (e.g. grasping, manipulating, tearing) and when it observes the same, or a similar, motor act performed by the experimenter or by a conspecific (Figure 15.3; Di Pellegrino et al., 1992; Gallese et al., 1996; Rizzolatti et al., 1996a). In contrast, they do not discharge during the observation of biological movements mimicking a motor act but devoid of a goal.

The two sets of mirror neurons most frequently found are those related to mouth and hand actions. Among mouth actions, biting and chewing are the most represented. The most common among hand actions are grasping, manipulating and holding. Although the visual responses of most mirror neurons are invariant with respect to visual aspects of the stimulus that trigger them, some of them show specificity for the direction of the hand movement, the space sector (left or right) in which the observed motor act is presented or the hand (left or right) used by the observed agent.

What is the functional role of the parieto-frontal mirror neurons? The most accepted hypothesis is that they are involved in understanding the goal of the observed motor acts. If this is correct,

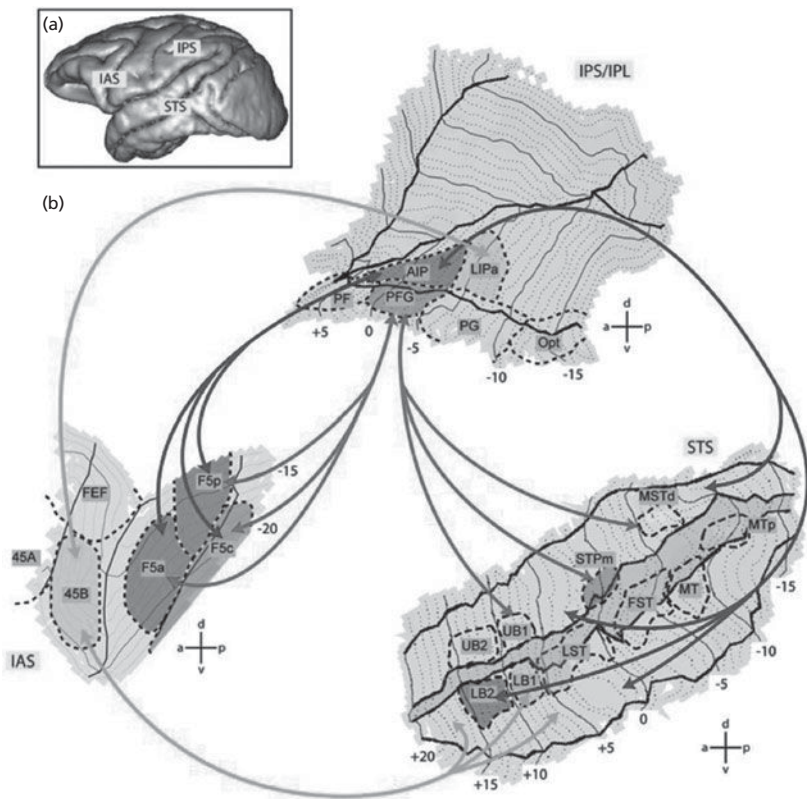


Figure 15.2 The mirror circuit for grasping. (a) The three main nodes of the mirror network. STS: superior temporal sulcus; IPS: intraparietal sulcus; IAS: inferior arcuate sulcus. The neurons located in the cortex within STS are visual neurons without mirror properties. Mirror neurons are located in the cortex within and around the other two other sulci. (b) Detailed organization of the grasping mirror circuit. Flattened representation of STS, IPS/IPL (inferior parietal lobule) and IAS. FEF = frontal eye fields, F5c = F5 convexity, F5p = F5 (bank) posterior, F5a = F5 (bank) anterior. Visual information on observed actions is sent forward from STS through parietal cortex to area F5 along two functional routes: a STPm—PFG—F5c and a LB2—AIP—F5a/p route, indicated with light and dark gray arrows, respectively. Area 45B receives parietal input from LIPa and also has direct connections with the lower bank STS (light gray arrows). The dark shaded areas specify the functional routes. Abbreviations: parietal and frontal lobe areas, see Fig. 1; STS fields are indicated with their conventional names: MT/V5, MTp, MSTd, FST, LST, STPm, lower bank 1 and 2 (LB1; LB2) upper bank 1 and 2 (UB1; UB2). See also Plate 7. Reproduced from Nelissen, K., Borra, E., Gerbella, M., Rozzi, S., Luppino, G., Vanduffel, W., Rizzolatti, G., & Orban, G.A., Action observation circuits in the macaque monkey cortex. *Journal of Neuroscience*, 31, 3743–3756 © 2011, The Society for Neuroscience, with permission.

they should become active, not only in response to visual stimuli, but also when there is sufficient information in the environment to understand the goal of the observed motor act.

Evidence in favor of this hypothesis came from two series of experiments: the first tested whether F5 mirror neurons were able to recognize actions from their sound (Kohler, Keysers, Umiltà, Fogassi, Gallese, & Rizzolatti, 2002), while the second tested whether the mental representation of an action could trigger their activity (Umiltà, Kohler, Gallese, Fogassi, Fadiga, Keysers, et al., 2001).

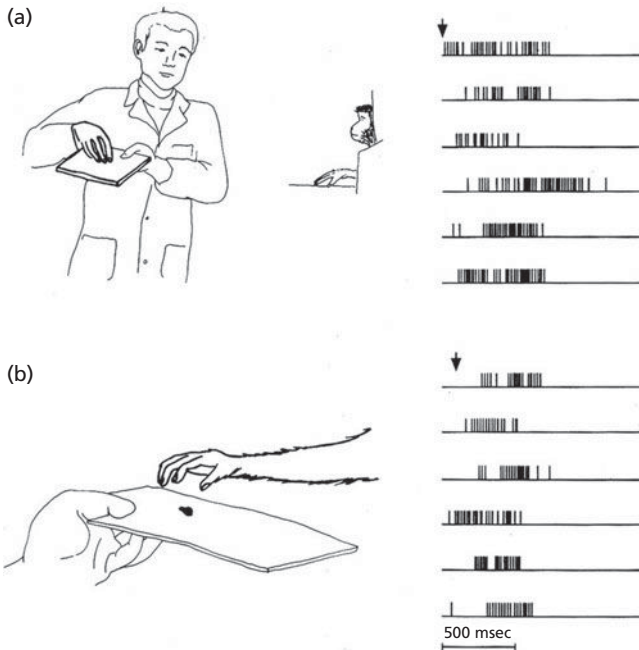


Figure 15.3 Example of a F5 mirror neuron. The neuron discharge during monkey grasping movements and when monkey observes grasping movements done by the experimenter.

Reproduced from di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V. and Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research* 91(1): 176–80. © 1992, with kind permission from Springer Science and Business Media.

Kohler et al. (2002) recorded F5 mirror neuron activity while the monkey was observing a “noisy” action (e.g. ripping a piece of paper), or was presented with the same noise without seeing it. The results showed that about 15% of mirror neurons responded to the presentation of actions accompanied by sounds and also responded to the presentation of the sound alone.

The rationale of the second experiment was the following. If mirror neurons are involved in action understanding, they should discharge also under conditions in which the monkey does not see the occurring action, but has sufficient clues to create a mental representation of what the experimenter does. The neurons were tested in two basic conditions. In one, the monkey was shown a fully visible action directed toward an object (“full vision” condition). In the other condition, the monkey saw the same action but with its final critical part hidden (“hidden” condition). Before each trial, the experimenter placed a piece of food behind the screen so that the monkey knew that there was an object there. Only those mirror neurons were studied that discharged to the observation of the final part of a grasping movement and/or to holding. The results showed that more than half of the tested neurons discharged in the hidden condition. Out of them, about half did not show any difference in the response strength between the hidden and full vision conditions, while the other half responded more strongly in the full vision condition.

In conclusion, both of these experiments showed that the activity of mirror neurons correlates with action understanding. The visual features of the observed actions are fundamental to trigger mirror neurons only inasmuch as they allow the understanding of the observed actions. If other

cues describing the action are available, mirror neurons signal the action even in the absence of visual stimuli.

Other properties of mirror neurons

A recent study investigated whether the discharge of mirror neurons of F5 is modulated by the distance at which the observed act is performed (Caggiano, Fogassi, Rizzolatti, Thier, & Casile, 2009). The results showed that this was indeed the case. About half of the recorded neurons showed stronger discharge when the action was in the monkey's peripersonal space (i.e. the space within its reach), while half preferred the extrapersonal space (i.e. the space outside its reach). Some of the studied neurons encoded space according to a metric representation. Most interestingly, other neurons encoded information in terms of operational space by discharging according to whether the monkey could interact with the object or not. These neurons may play an important role in deciding the monkey's future action.

Typically, the visual responses of mirror neurons have been studied in a naturalistic context. More recently, however, in order to assess the mirror neuron properties in a more quantitative way, the discharge of mirror neurons has been recorded while the monkey observed video clips showing different motor acts. It was found that mirror neurons do not require necessarily the observation of a live person interacting with objects. They also respond to actions shown in movies, although their discharge is typically weaker in this last condition.

Using filmed motor acts Caggiano, Fogassi, Rizzolatti, Pomper, Thier, Giese, et al. (2011) studied the responses of mirror neurons to the observation of motor acts performed in an egocentric (subjective) perspective or in two types of third-person views (lateral or frontal view). The results showed that the responses of about 25% of the tested neurons was invariant relative to the point of view from which the motor act was observed, while the remaining ones appeared to encode the three different perspectives in equal percentage. Examples are shown in Figure 15.4.

The authors proposed that view-dependent mirror neurons might provide, through feedback connections, a link between neurons coding the goal of an observed motor act and neurons that code its pictorial aspects. In this way, F5 neurons, through a bottom-up mechanism, could determine a full perception of the observed motor act, which includes its meaning (goal) as well as the its visual details.

A recent study by Kraskov, Dancause, Quallo, Shepherd, & Lemon (2009), showed that about half of the pyramidal tract neurons (PTNs) recorded from area F5, exhibited mirror activity. Furthermore, some pyramidal tract neurons fired actively when the monkey grasped food, but showed suppression during action observation. The authors suggested that PTNs might play a role in inhibition of self-movements during action observation.

It was recently shown that there is another area of the parietal lobe that contains neurons with mirror properties—area VIP (Ishida, Nakajima, Inase, & Murata, 2009). This area, buried in the intraparietal sulcus, forms with frontal area F4 a circuit transforming somatosensory sensory and visual stimuli presented in the space around the monkey body-parts (peripersonal space) into head, mouth and arm movements (Colby, Duhamel, & Goldberg, 1993; Duhamel, Colby, & Goldberg, 1998; Fogassi, Gallese, Fadiga, Luppino, Matelli, & Rizzolatti, 1996; Gentilucci, Scandolara, Pigarev, & Rizzolatti, 1983; Graziano, Yap, & Gross, 1994).

Ishida and colleagues demonstrated that many VIP neurons, coding peripersonal space, also respond to stimuli presented in the peripersonal space of another individual located at about one meter far from the monkey and facing it. It is important to keep in mind that, although Ishida and colleagues did not study motor responses, area VIP is strictly connected with area F4. It is plausible,

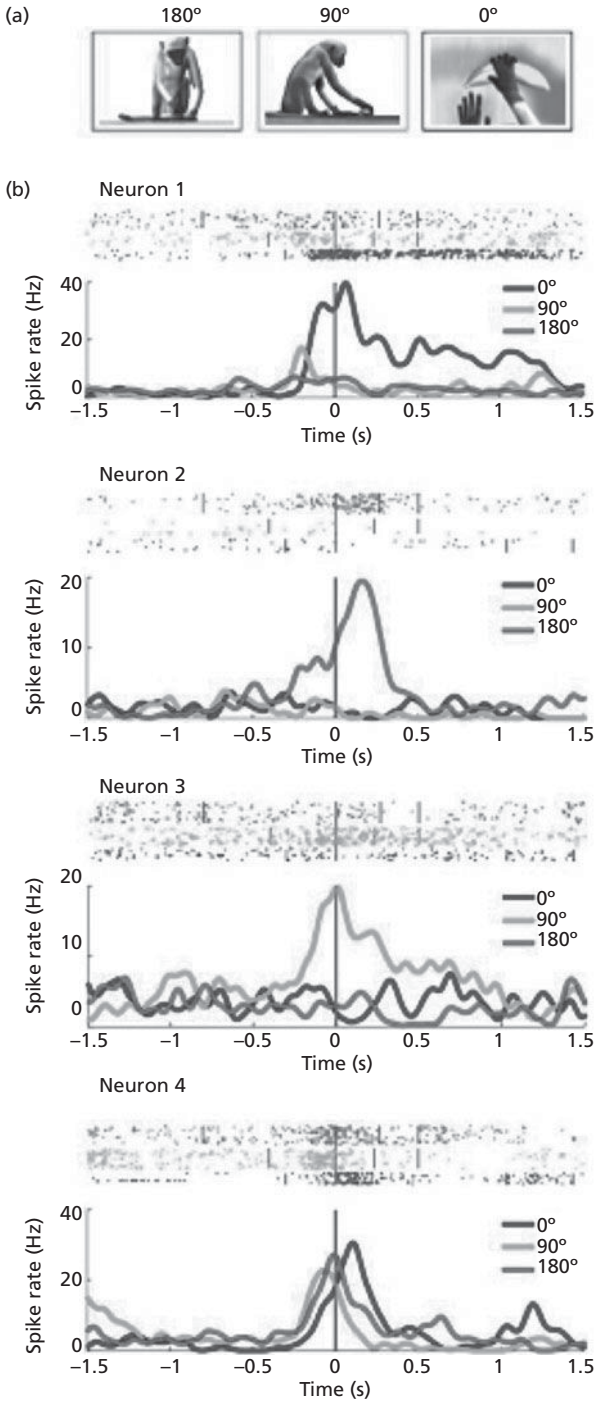


Figure 15.4 Responses of mirror neurons to actions observed from three different points of view. (a) Experimental conditions (frontal point of view: 180°; side view: 90°; subjective point of view: 0°); (b) Responses of four mirror neurons during observation of filmed goal-directed actions. Responses were aligned with the moment when the monkey touched the object). Neuron 1 showed a selectivity for actions presented in the subjective point of view (0°). Neuron 2 showed a modulation for actions presented from a frontal point of view (180°), Neuron 3 showed a modulation for actions presented from a side view (90°). The activity of Neuron 4 was modulated by actions seen from all tested points of view. See also Plate 8.

Reproduced from Caggiano, V., Fogassi, L., Rizzolatti, G., Pomper, J. K., Thier, P., Giese, M. & Casile, A. View-based encoding of actions in mirror neurons of area F5 in macaque premotor cortex. *Current Biology* 21(2): 144–8.

© 2011, with permission from Elsevier.

therefore, that neuronal responses that seem merely visual actually represent potential motor acts directed toward specific body parts.

The study of VIP neurons is of great interest because it shows that the mirror mechanism of this area encodes body-directed, rather than object-directed motor acts, thus opening fascinating possibilities on the mechanisms for encoding the body of others.

There are no single neuron studies that investigated the presence of mirror neurons in area F4 and in the ventro-rostral sector of F2, sectors that also receive visual input from posterior parietal lobe. Human studies (Filimon, Nelson, Hagler, & Sereno, 2007, see also Buccino, Binkofski, Fink, Fadiga, Fogassi, Gallese, et al., 2001) suggest, however, that reaching mirror neurons could be present in these areas.

One must be very cautious in considering as “mirror area” a motor area that responds to visual stimuli without an extremely accurate control of the possibility of unspecific motor activity during visual stimulus presentation. This is particularly true for the superior parietal lobule, dorsal premotor cortex and primary motor cortex, all areas that are involved in covert motor preparation (Crammond & Kalaska, 2000; Kalaska & Crammond, 1995), besides motor execution. Thus, the findings of visual responses in these areas during action observation obtained in 2-DG experiments (e.g. Raos, Evangelidou, & Savaki, 2007) should be considered as only indicative of the possibility of the presence of mirror neurons in those areas.

Functional proprieties of mirror neurons: understanding the intentions of others' motor acts

Single neuron studies show that mirror neurons are not only crucial for understanding motor acts, but can also be involved in coding the motor intention of an action performed by another individual (Bonini, Rozzi, Serventi, Simone, Ferrari, & Fogassi, 2009; Fogassi et al., 2005).

In order to understand the further properties of mirror neurons, it is necessary to clearly define the difference between “motor act” and “action.” We define motor act as a series of movements with a specific goal. We define action as a series of motor acts that, when executed, allow the achievement of biological useful goals.

In order to investigate the neural basis of motor intention coding, experiments were carried out in which we assessed whether neurons discharging during the execution of grasping were influenced by the type of action in which they were embedded (Bonini et al., 2009; Fogassi et al., 2005). Grasping neurons were recorded from areas PFG and F5 while the monkey executed a motor task or observed the same task (performed by an experimenter) in which the motor act (grasping) was embedded into two different actions (grasping to eat and grasping to place) (Figure 15.5). The results showed that a high percentage of parietal and premotor neurons discharged differently during the execution of grasping, depending on the final goal of the action (either eating or placing) in which grasping was embedded. On the basis of these findings it has been proposed (Fogassi et al., 2005; Rizzolatti, Ferrari, Rozzi, & Fogassi, 2006) that parietal and premotor neurons form pre-wired chains, in which a neuron coding a given motor act is facilitated by the previously executed one. According to this proposal, a neuron coding for instance a grasping act, is not a multi-purpose unit, but belongs to a chain aimed for a specific ultimate goal. Thus, any time an agent has the intention (final goal) of performing an action, specific neuronal chains are activated. Note that this model would also account for the fluidity with which the different motor acts of an action are executed one after another, as shown by kinematic studies (Jeannerod 1988; Rosenbaum, Cohen, Jax, Weiss, & van der Wel, 2007).

Similarly to the motor task, during the visual task it has been found that most mirror neurons discharge differently when observing a grasping act embedded within different actions (Figure 15.5).

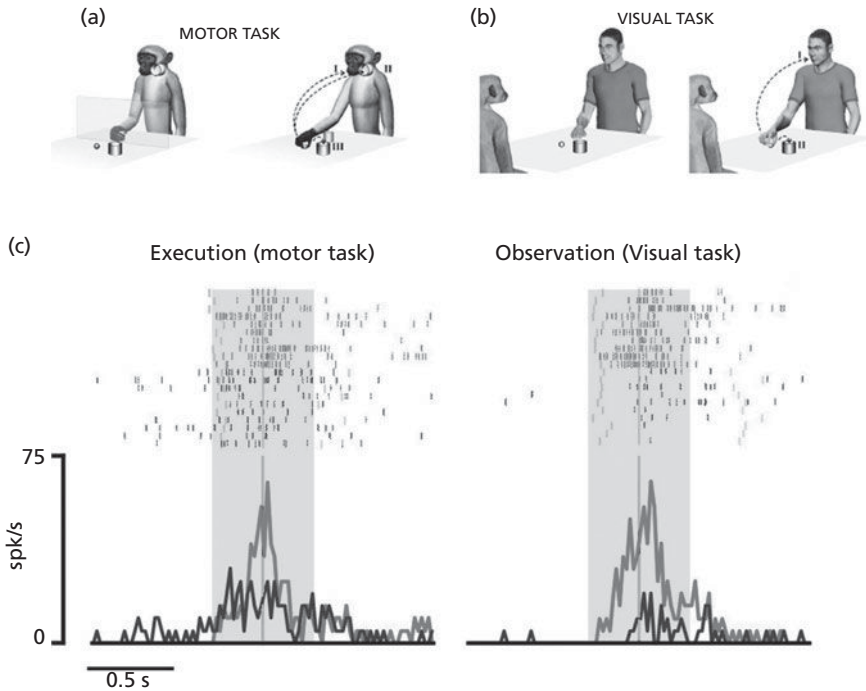


Figure 15.5 Example of an F5 mirror neuron modulated by action intention. (a) Paradigm used for the motor task. The monkey, starting from a fixed position, reaches and grasps a piece of food or an object, then it brings the food to the mouth and eats it (I, grasp-to-eat), or places it into a container (II/III, grasp-to-place). (b) Paradigm used for the visual task. The experimenter, starting from a fixed position, reaches and grasps a piece of food or an object, then brings the food to the mouth and eats it (I, grasp-to-eat) or places it into a container (II, grasp-to-place). (c) Discharge of the neuron during execution (left) and observation (right) of the two actions. Rasters and histograms are aligned (green lines) on the contact between the monkey or experimenter's hand and the object. Red: neuron discharge during grasp-to-eat condition; gray: neuron discharge during grasp-to-place condition. Blue bars indicate the onset of the hand movement, yellow bars indicate the contact between the hand and the container in the grasp-to-place condition. See also Plate 9.

Adapted from Bonini, L., Rozzi, S., Serventi, F. U., Simone, L., Ferrari, P. F., & Fogassi, L. Ventral premotor and inferior parietal cortices make contribution to action organization and intention understanding. *Cerebral Cortex* 20(6): 1372–85. © 2010, Oxford University Press, with permission. For permission to reuse this material, please visit <http://www.oup.co.uk/academic/rights/permissions>.

It was proposed that the neuronal selectivity for the action goal during grasping observation represents a prediction of the action outcome. Thus, in agreement with the “chain” interpretation of the results of the motor task, the observation of a motor act embedded in an action, would activate a chain corresponding to a specific intention. This activation would allow one to understand immediately the motor intentions of others.

The mirror mechanism in humans

Immediately following the discovery of mirror neurons in the monkey, a large number of neurophysiological and brain-imaging studies demonstrated the presence of a mirroring mechanism

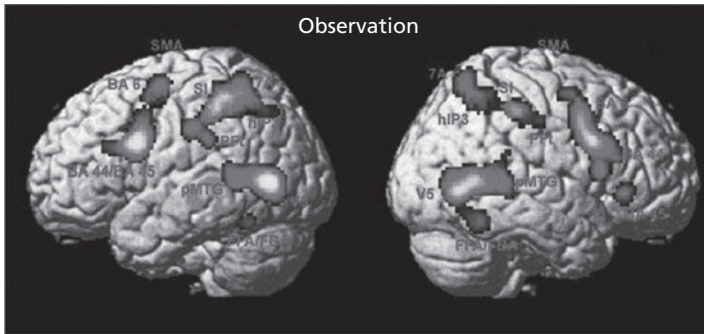


Figure 15.6 Lateral view of the two hemispheres of the human brain showing areas activated in neuroimaging studies during action observation. The depicted activations result from a meta-analysis performed on 87 studies. The three main nodes of activation are: the superior temporal sulcus; the inferior parietal lobule, including the intraparietal sulcus plus a small part of superior parietal lobule; the premotor cortex, mainly its ventral part, plus the posterior part of the inferior frontal gyrus. See also Plate 10.

Adapted from Caspers, S., Zilles, K., Laird, A. R., & Eickhoff, S. B. ALE meta-analysis of action observation and imitation in the human brain. *NeuroImage* 50(3): 1148–67. © 2010, with permission from Elsevier.

in humans and described its basic properties (see Fabbri-Destro & Rizzolatti, 2008; Rizzolatti & Craighero, 2004). In the following sections we will review the most significant studies on the functional role of the mirror mechanism in human behavior.

fMRI studies: the localization of the human mirror mechanism

Initial PET studies already showed that the mirror mechanism for hand-grasping actions was located in humans in the same regions as in the monkey. These regions comprised IPL, including the cortex located inside the intraparietal sulcus, and the ventral sector of the precentral gyrus plus the caudal part of the inferior frontal gyrus (Grafton, Arbib, Fadiga, & Rizzolatti, 1996; Rizzolatti, Fadiga, Matelli, Bettinardi, Paulesu, Perani, et al., 1996b). These data were fully confirmed in the following years by a large number of fMRI studies (Caspers, Zilles, Laird, & Eickhoff, 2010; see Figure 15.6).

A series of studies of cortical activations during the observation of transitive motor acts showed that, in the precentral gyrus, these acts are localized according to a rough somatotopic organization similar to that of the classical homunculi of motor physiology (Buccino et al., 2001; Sakreida, Schubotz, Wolfensteller, & von Cramon, 2005; Saygin, Wilson, Hagler, Bates, & Sereno, 2004; Shmuelof and Zohary, 2006; Ulloa & Pineda 2007; Wheaton, Carpenter, Mizelle, & Forrester, 2008). The observation of motor acts done by mouth, hand, and foot, produce a somatotopic activation: the leg motor acts were located dorsally, the mouth ventrally, and the hand in an intermediate position.

A somatotopic organization is also present in the inferior parietal lobe along to and within the intraparietal sulcus. The mouth is located rostrally, the hand in an intermediate position and the leg caudally (Buccino et al., 2001). A recent study attempted to better define this organization (Jastorff, Begliomini, Fabbri-Destro, Rizzolatti, & Orban, 2010). Four motor acts (grasping, dragging, dropping, and pushing) performed with the mouth, hand and foot were presented to volunteers. The results showed that, superimposed to the somatotopic organization, the parietal

organization appears to be also influenced by the behavioral significance of the observed motor acts. For the motor acts tested in that study the activation was mostly found in area AIP, that is an area coding hand movements. Furthermore, there was a subdivision between the localization of self-directed (grasping and dragging) and outward-directed motor acts (dropping and pushing). The former activated ventral sectors of AIP, while the latter activated the dorsal sectors of the same region, as well as the adjacent dorso-caudal cortex. It therefore appears that in the premotor cortex, acts that are executed with the same effector tends to be clustered together, while the parietal cortex tends to classify the motor acts on the basis of their positive or negative meaning for the observer.

Some experiments addressed the issue of the localization of the observation of reaching movements. The cortical representation of reaching was studied by Filimon et al. (2007) in a fMRI study. They used as stimuli photographs of abstract wooden shapes. These abstract shapes served as target for the reach. They were chosen in order to prevent object verbalizing and to motivate the reach because naturalistic reaching usually occurs toward objects. The results showed activation in the superior parietal and dorsal premotor cortex (see also Buccino et al., 2001). These data suggest that reaching mirror neurons could be present in these areas.

The dorsal premotor cortex and superior parietal lobule have also been occasionally found to be active during the observation and execution of grasping movement (Gazzola & Keysers, 2009; Grèzes, et al., 2003). Although it is possible that these activations are due to a mirror mechanism, it is more likely that they reflect motor preparation. In favor of this interpretation are monkey single-neuron data showing the involvement of these areas in covert motor preparation (Kalaska & Crammond, 1995; Crammond & Kalaska, 2000).

The mirror mechanism is involved in understanding the goal of others' motor acts

Various fMRI studies provided evidence that, as in the monkeys, the human mirror mechanism located in the parieto-frontal circuits is involved in understanding the goal of the observed motor acts. Gazzola and colleagues (Gazzola, Rizzolatti, Wicker, & Keysers, 2007) instructed volunteers to observe video-clips in which either a human or a robot arm grasped objects. In spite of differences in shape and kinematics between the human and robot arms, the parieto-frontal mirror circuit was activated in both conditions.

These results were extended by Peeters and colleagues (Peeters, Simone, Nelissen, Fabbri-Destro, Vanduffel, Rizzolatti, et al., 2009), who investigated the cortical activations in both monkeys and humans in response to the observation of motor acts performed by a human hand, a robot hand, and a variety of tools. They found that regardless of the type of effector used, the hand-grasping network, formed by intraparietal and ventral premotor cortex, was always active in both humans and monkeys. Interestingly, only in humans, the observation of tool motor acts produced activation in a rostral sector of the left anterior supramarginal gyrus, which was not activated during the observation of hand grasping movements. This finding and the lack of a region in the monkey specifically active during tool act observation even following prolonged tool use, suggest that, in evolution, humans acquired a specific brain region devoted to tool use and tool use understanding.

While most studies on mirror networks used visual stimuli, some presented auditory stimuli, contrasting those typical of specific actions with others possessing characteristics unrelated to human actions. In an fMRI study, Lewis, Brefczynski, Phinney, Janik, & DeYoe (2005) presented to volunteers animal vocalization and sounds of tools manipulated by hands. The results showed a clear dissociation between the two types of stimuli. Hearing and categorizing animal vocalizations preferentially activated the middle portion of the superior temporal gyrus bilaterally, while

hearing and categorizing sounds of tools activated the parieto-frontal mirror circuit. Similarly, Gazzola, Aziz-Zadeh, & Keysers (2006) showed that listening to the sound of hand and mouth motor acts activated this last circuit and the activation was somatotopically organized in the left premotor cortex.

There is no doubt, that, in some cases, the observation of motor behavior might require a mechanism different from mirroring in order to be understood. The capacity of humans to recognize animals' actions that do not belong to human motor repertoire and cannot be captured by a motor generalization, is a typical example in this regard. Evidence for a non-mirror mechanism in action recognition has been provided by Buccino and colleagues (2004a) in an fMRI study in which volunteers were presented with video clips showing motor acts that did (biting and reading) or did not belong to the human motor repertoire (lip smacking and barking). Although all volunteers recognized the observed motor acts, no activation of parieto-frontal mirror areas was found for those actions that did not belong to human motor repertoire (e.g. barking). The areas that became active in the last case were occipital and STS areas. By contrast, the sight of motor acts that were within the observers' motor repertoire (e.g. dog, monkey and human biting) activated the parieto-frontal mirror network.

These data indicate that the recognition of others' motor behavior can rely, in some instances, on the mere processing of its visual aspects. This visual recognition appears similar to that carried out in the ventral stream areas for the recognition of inanimate objects. It allows for the recognition of the observed behavior, but does not provide the observer with cues that are necessary for a real understanding of the conveyed message (e.g. the communicative intent of the barking dog). By contrast, when the observed action "intrudes" into the motor system through the mirror mechanism, that action is not only visually recognized but also understood, because its motor goal-relatedness is shared by the observer and the agent. In other words, the observed action is understood from the inside as a motor possibility and not just from the outside as a mere visual experience.

This point was recently discussed by Frith & Frith (2007). They maintained that, although the parieto-frontal mirror mechanism was active in all conditions in which the motor task has to be directly understood, when volunteers were required to judge the reasons behind the observed actions there was an activation of a sector of the anterior cingulate cortex and of other areas of the so-called "mentalizing network" (e.g. STSp and tempo-parietal junction; see Frith & Frith 2007). Activation of the same network was also shown in a study that investigated unusual actions performed in implausible vs. plausible contexts, as well as in experiments that studied the neural basis of reason inference in non-stereotypic actions (Brass, Schmitt, Spengler, & Gergely, 2007; de Lange, Spronk, Willems, Toni, & Bekkering, 2008; Liepelt, Von Cramon, & Brass, 2008).

TMS studies show that, in humans, both transitive and intransitive gestures are coded by the mirror mechanism

Crucial evidence that the motor system in humans has mirror properties was provided by transcranial magnetic stimulation (TMS) studies. Fadiga et al. (1995) recorded motor-evoked potentials (MEPs), elicited by stimulation of the left motor cortex, from the right hand and arm muscles in volunteers required to observe an experimenter grasping objects (transitive hand actions) or performing meaningless arm gestures (intransitive arm movements). The results showed that the observation of both transitive and intransitive actions determined an increase of the recorded MEPs with respect to the control conditions. The increased selectively concerned those muscles that participants used for producing the observed movements.

Strafella & Paus (2000) used a double-pulse TMS technique in order to establish whether the duration of intracortical recurrent inhibition observed during action execution, also occurred during action observation. The results confirmed this hypothesis suggesting that MEP's facilitation during movement observation results from a facilitation of the primary motor cortex, due to mirror activity of the premotor areas, rather than from other mechanisms.

The finding that, in humans, intransitive movements, and not only goal-directed actions, are coded by the mirror mechanism was confirmed by Maeda et al. (2002) in a study in which they investigated the effect of hand orientation on cortical excitability. In a fundamental study, Gangitano et al. (2001) recorded MEPs from the hand muscles of normal volunteers, while they were observing grasping movements made by another individual. The MEPs were recorded at different intervals following the movement onset. The results showed that the motor cortical excitability faithfully followed the grasping movement phases of the observed action. This represents a crucial demonstration of the direct coding (not semantically mediated) of the observed motor acts.

Recently Cattaneo and colleagues (Cattaneo et al., 2009) showed that mirror coding might depend on the content of the observed behavior. They recorded motor-evoked potentials (MEPs) to TMS from the right *opponens pollicis* (OP) muscle in participants that observed an experimenter either merely opening and closing normal and reverse pliers or using them to grasp objects. The observation of tool movements (i.e. opening and closing the pliers without grasping anything) activated a cortical representation of the hand movements involved in the observed motor behavior. By contrast, the observation of the tool grasping action activated a cortical representation of the observed motor goal, irrespective of the individual movements and the order of movements required to achieve it. These findings crucially demonstrate that the human parieto-frontal mirror network encodes not only motor acts, but also movements.

Encephalographic recordings studies confirm that observed motor acts are directly coded on the motor cortical system

Movement execution determines a desynchronization of the rhythms recorded from the rolandic (motor) region. A typical rolandic rhythm is “mu” rhythm. This rhythm first described by Gastaut et al. (1952) under the name of “rolandic rhythm *en arceau*”, has a frequency around 13 Hz and a particular arch-like morphology. Subsequent analysis (Hari, 2006; Tiihonen et al., 1989) revealed that its arch-like appearance is due to the coexistence of (at least) two not harmonic frequency components whose spectral peaks were distributed around 10 Hz (alpha band) and 20 Hz (beta band).

Altschuler and co-workers (Altschuler et al., 1997) reported that, besides execution, desynchronization of mu rhythm can also be achieved by observing movements. This finding was later confirmed by several researchers (Babiloni et al., 2002; Caetano et al., 2007; Cochin et al., 1998, 1999; Muthukumaraswamy & Johnson, 2004; Muthukumaraswamy et al., 2004; Nishitani & Hari, 2000; Perry & Bentin, 2009; Press et al., 2011). It was, therefore, proposed, that the cortical desynchronization during action observation is due to the mirror mechanism—observing others directly activates the premotor cortex and consequently it's connected motor areas. This determines the occurrence of the rolandic desynchronization.

This desynchronization explanation was confirmed by Nishitani & Hari (2000). Using MEG, they asked participants to either grasp an object with their right hand, observe the same action being performed by the experimenter, and to observe and replicate the seen action. The results showed that, during execution of the motor act, there was an initial activation of the left inferior frontal cortex (area 44) followed by that of the left primary motor cortex. During observation

and imitation, the sequence of activations was similar although obviously starting from the visual areas.

Further studies investigated if there is a dynamic correlation between the desynchronization of rolandic rhythms during movement execution, movement imagery, and movement observation. The data showed that the velocity of both executed and imagined movement modulates the same cortical frequency bands in a similar fashion (Kilner et al., 2003; Press et al., 2011; Yuan, Perdoni, & He, 2010) reported that a similar modulation also occurs during the observation of a biological motion of the arm. These findings provide further strong evidence for a direct (not semantically mediated) mapping mechanism of the observed movements on the motor cortex (the mirror mechanism).

Single mirror neuron data in humans

Single neurons recordings in humans are limited by obvious ethical reasons. The same reasons limit the localization of cortical brain regions from which single neurons can be recorded. Thus, there are no available data, up to now, on mirror neurons located in the parieto-frontal circuits. In contrast, neurons with apparent mirror properties have been recorded from human medial frontal and temporal cortices in epileptic patients. These neurons were tested, while patients executed or observed hand grasping actions and facial emotional expressions (Mukamel et al., 2010).

Many of these neurons, located in the supplementary motor areas and the hippocampus, responded to both observation and execution of these actions. Others showed excitation during action execution and inhibition during action observation. The precise functional meaning of these neurons is not clear and, unfortunately, there are no monkey data that may throw light on their behavioral significance. However, judging from the type of structures from which they have been recorded, it may be possible that mirroring could be also involved in cognitive functions higher than those discussed in the present chapter.

Mirror neurons have also been recorded from a sector of the anterior cingulate cortex processing sensory information for pain. This type of neuron responded both to thermal stimuli in the noxious range and to observation of painful stimuli delivered to the examiner, although, in this case, with less intensity (Hutchinson et al., 1999).

The mirror mechanism is involved in imitation and imitation learning

The term imitation has different meaning in various research fields (see Hurley & Chater, 2005). Typically, experimental psychologists use this term to define the capacity of an individual to replicate an observed motor act after having seen it executed by others. Ethologists define imitation as the capacity to acquire a motor behavior previously not present in the observer's motor repertoire by observation, and to replicate it using the same movements employed by the teacher (Tomasello & Call, 1997).

Imitation as a replica of a motor act already present in observer's motor repertoire has been extensively investigated by Prinz and his coworkers (Prinz, 2002). They established that the more a motor act resembles one that is present in the observer's motor repertoire, the greater the tendency to do it. Perception and execution must therefore possess, according to them, a "common representational domain."

The discovery of mirror neurons suggested a reformulation of this concept by considering the "common representational domain" not as an abstract, amodal domain (Prinz, 1987), but rather as a motor mechanism directly activated by the observed actions.

It was argued against this proposal that monkeys are unable to imitate and, therefore, the mirror mechanism cannot be involved in this capacity. However, as already mentioned, the human mirror mechanism is able to code intransitive meaningless gestures. Direct evidence that the human mirror mechanism is involved in imitation was provided by an fMRI study by Iacoboni et al. (1999). These authors studied normal human volunteers in two conditions: “observation-only” and “observation-execution”. In the “observation-only” condition, volunteers were shown a moving finger, a cross on a stationary finger, or a cross on empty background. The instruction was to observe the stimuli. In the “observation-execution” condition, the same stimuli were presented, but this time, the instruction was to lift the right finger, as fast as possible, in response to them. The crucial contrast was between the trials in which the volunteers made the movement in response to an observed action (“imitation”) and the trials in which the movement was triggered by the cross (a non-imitative behavior). The results showed that the activation was stronger during “imitation” in the posterior part of the inferior frontal gyrus. Similar results were also obtained by Koski et al. (2002) and Grèzes et al. (2003).

Further evidence that the mirror mechanism is involved in imitation was found using repetitive TMS (rTMS). In a group of volunteers the caudal part of the left inferior frontal gyrus (Broca’s area) was stimulated while they (a) pressed keys on a keyboard, (b) pressed the keys in response to a point of red light which, directed onto the keyboard, indicated which key to press, or (c) imitated a similar movement executed by another individual. The data showed that rTMS lowered the participants’ performance during imitation, but not during the other two tasks (Heiser, Iacoboni, Maeda, Marcus, & Mazziotta, 2003).

The mirror mechanism is also involved in imitation learning. An interesting model based on ethological studies of ape behavior has been proposed by Byrne (2002). According to this model, learning by imitation results from the integration of two distinct processes. First, the observer segments the action to be imitated into its individual elements, thus converting it into a string of acts belonging to the observer motor repertoire. Secondly, the observer organizes these motor acts into a sequence that replicates that of the demonstrator.

The neural basis of imitation learning was investigated by Buccino et al. (2004b) in an fMRI study. Musically naive participants were asked to imitate guitar chords played by an expert guitarist. Cortical activations were mapped during the following events: (a) observation of the chords made by the expert player, (b) pause, (c) execution of the observed chords, and (d) rest. In addition, there were other conditions to control for observation not followed by imitation and for non-imitative motor activity. The results showed that during observation for imitation there was activation of a cortical network formed by IPL and the dorsal part of PMv, plus the *pars opercularis* of IFG. This circuit was also active during observation in the control conditions in which participants merely observed the chords, or observed them with the instruction to subsequently perform an action not related to guitar chord execution. During the pause, activation was found in the imitation condition in the same circuit as during observation, but, most interestingly, also in the middle frontal cortex (area 46) and in the anterior mesial cortex. Motor activations dominate the picture during chord execution.

A subsequent fMRI study, based on a similar experimental design, but carried out in both expert and naive guitarists confirmed the joint role of the mirror areas and prefrontal lobe in imitation learning. In particular, it was confirmed the fundamental role that area 46 plays in combining different motor acts in a new specific motor pattern (Vogt, Buccino, Wohlschläger, Canessa, Shah, & Zilles, 2007).

In summary, these data show that during new motor pattern formation there is a strong activation of the mirror mechanism. Imitation learning is, however, based on a two-step mechanism. First,

the observed actions are decomposed into elementary motor acts that activate, via mirror mechanisms, the corresponding motor representations in the parietal and frontal lobe. Subsequently, these motor representations are re-combined to fit the observed model. For this re-combination, a crucial role is played by frontal area 46.

Emotions and the mirror mechanism

According to Ekman (1999), humans share five basic emotions: fear, sadness, happiness, anger, and disgust. This last emotion has been classified among the basic ones also by Darwin (1872).

Disgust is one of the emotions most investigated in neurophysiological studies. Brain imaging studies showed that when an individual is exposed to disgusting odors or tastes, there is an intense activation of two structures: the amygdala and the insula (Augustine, 1996; Royet, Plailly, Delon-Martin, Kareken, & Segebarth, 2003; Wicker, Keysers, Plailly, Royet, Gallese, & Rizzolatti, 2003).

The insula is a complex structure. A recent meta-analysis by Kurth, Zilles, Fox, Laird, & Eickhoff (2010), based on a large number of fMRI studies, provided a comprehensive correlative functional picture of human insula. According to Kurth et al. there are four distinct functional fields in human insula: the sensorimotor, the socio-emotional, the olfactory-gustatory and, finally, the cognitive field. A conjunction analysis across these domains revealed that, aside from the sensorimotor field, all the other ones share activations in the cognitive field, a sector located in the anterior dorsal insula.

The sensorimotor field is located in the dorsal-posterior part of the insula. It is adjacent to SII and appears to represent a ventral extension of the primary somatosensory cortex. This field mediates elaboration of sensory information similar to that carried out in SII and adjacent areas. It is **not** related to emotions. The anterior sector receives a rich input from olfactory and gustatory centers and, in addition, a visual input from the inferotemporal lobe (Mesulam & Mufson, 1982). Its electrical stimulation produces motor behaviors related to food ingestion (dorsal part) and disgust responses (ventral part). The part of the insula concerned with emotional behaviors is located in the ventral part of the insula. In both humans and monkeys, the electrical stimulation of the insula produces body movements often accompanied, in its ventral part, by autonomic and visceromotor responses (Caruana, Jezzini, Sbriscia-Fioretti, Rizzolatti, & Gallese, 2011; Krolak-Salmon, Henaff, Isnard, Tallon-Baudry, Guenot, Vighetto, et al., 2003; Penfield & Faulk, 1955).

On the basis of previous brain imaging studies (Phillips, Young, Scott, Calder, Andrew, Giampietro, et al., 1998; Sprengelmeyer, Rausch, Eysel, & Przuntek, 1998; Zald, & Pardo, 2000) showing that, in humans, observation of faces showing disgust activates the anterior insula, Wicker et al. (2003) investigated whether the area within the insula that activates during the **experience** of disgust, would also show activation during the observation of faces expressing disgust.

This fMRI study consisted of two sessions. In the first, the participants were exposed to unpleasant and pleasant odorants; in the second they watched a video showing the face expression of people sniffing an unpleasant, a pleasant or a neutral odor. Three main structures became active during the exposure to smells: the amygdala, the insula and the anterior cingulate. The amygdala was activated by both unpleasant and pleasant odors. In the insula, pleasant odorants produced a relatively weak activation located in a posterior part of the right insula, while disgusting odorants activated the anterior sector bilaterally. The results of visual runs showed activations in various cortical and subcortical centers, but not in the amygdala. The left anterior insula and the anterior cingulate were activated only during the observation of disgust.

The most important result of the study was the demonstration that precisely the same foci within the anterior insula that were activated by the exposure to disgusting odorants were also activated by the observation of disgust. These data strongly suggest that the insula (and the anterior cingulate) contain neural populations that becomes active both when the participants experience disgust and when they see it in others.

The hypothesis that we recognize others' emotion by activating structures mediating the same emotion in ourselves has been advanced by various authors (Calder, Keane, Manes, Antoun, & Young, 2000; Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003; Gallese, Keysers, & Rizzolatti, 2004; Singer, Seymour, O'Doherty, Kaube, Dolan, & Frith, 2004). Particularly influential in this respect has been the work by Damasio (2003). According to his findings, mostly based on brain lesions, the neural basis of emotion understanding is the activation of an "as-if-loop", the core structure of which is the insula (Damasio, 2003).

It should be stressed that the direct activation of visceromotor structures like the insula does not exclude that we may recognize emotions indirectly, using cognition. Some particular visual feature representing emotion, like in schematic faces, can determine emotion recognition without necessarily eliciting the observed emotion.

The mirror mechanism and autism

Autistic syndrome disorder (ASD) is a heterogeneous developmental syndrome characterized by marked impairments in social interaction and communication as well as increased repetitive stereotyped behaviors and/or restricted interests (Kanner, 1943).

Although ASD is not associated with severe motor disturbances, many studies have reported motor deficits including alterations in motor milestone development (Teitelbaum, Teitelbaum, Nye, Fryman, & Maurer, 1998), clumsiness, motor incoordination, disturbances in reach-to-grasp movement (Ghaziuddin & Butler, 1998; Miyahara, Tsujii, Hori, Nakanishi, Kageyama, & Sugiyama, 1997), deficits in gross and fine motor movement (Noterdaeme, Mildenberger, Minow, & Amorosa, 2002), and impaired postural control (Kohen-Raz, Volkmar, & Cohen, 1992; Minshew, Sung, Jones, & Furman, 2004). These disturbances are to be added to those dealing with intrusive and abnormal movements, including repetitive hand flapping, stereotypy, and self-injurious behaviors (Gritti, Bove, Di Sarno, D'Addio, Chiapparò, & Bove, 2003; Mooney, Gray, & Tonge, 2006; Nayate, Bradshaw, & Rinehart, 2005). Furthermore, there is preliminary evidence that praxis deficits in children with ASD correlate positively with social, communicative, and behavioral impairments (Dowell, Mahone, & Mostofsky, 2009; Dziuk, Gidley Larson, Apostu, Mahone, Denckla, & Mostofsky, 2007; Qiu et al., 2010).

Taken together, this impressive amount of data indicating a profound deficit in the motor system organization, suggest, a priori, the existence of a deficiency of a normal development of the mirror mechanism, which is a motor mechanism, in ASD. Note also that the crucial symptoms of ASD (impairment in communication, language, and emotion) appear to match functions that are mediated by the mirror mechanism. On the basis of these considerations the hypothesis has been advanced (Williams, Whiten, Suddendorf, & Perrett, 2001; see Ramachandran 2011) that there is a relation between a deficiency in the abilities underpinned by mirror mechanism and the core deficit of autism.

Some evidence in favor of this hypothesis comes from EEG studies on mu rhythm in autistic children. They showed that while individuals with autism present a suppression of mu rhythm during voluntary movements, this suppression is absent when they watched someone else performing the movement (Ramachandran 2011). Similar data were also found by Martineau and

colleagues (Martineau, Cochin, Magne, & Barthelemy, 2008). Oberman, Ramachandran, & Pineda (2008) investigated how familiarity between the observing individual and the person performing the movements modulated the entity of mu rhythm suppression. Their study revealed that mu suppression depended on the familiarity of the observer with the agent and that children with autism might show mu suppression when a familiar person performs the action.

Evidence in favor of a deficit of the mirror mechanism in ASD also came from an fMRI study. High functioning children with autism and matched controls were scanned while imitating and observing emotional expressions. The results showed a significantly weaker activation in inferior frontal gyrus (IFG) in children with autism than in typically developing (TD) children. Most interestingly, the activation was inversely related to symptom severity (Dapretto, Davies, Pfeifer, Scott, Sigman, Bookheimer, et al., 2006).

Impaired motor facilitation during action observation was also reported in autism using TMS (Enticott, Kennedy, Rinehart, Tonge, Bradshaw, Taffe, et al., 2012; Théoret, Halligan, Kobayashi, Fregni, Tager-Flusberg, & Pascual-Leone, 2005). Furthermore, unlike TD individuals that, when viewing persons face-to-face, tend to imitate them in a mirror way, children with autism do not show this preference (Avikainen, Wohlschläger, Liuhanen, Hänninen, & Hari, 2003). This imitation peculiarity is most likely due to a deficit of mirror mechanism coding other person's movements on one's own.

Recently, the deficit of the mirror mechanism in autism has been addressed from another perspective (Cattaneo, Fabbri-Destro, Boria, Pieraccini, Monti, Cossu et al., 2007). TD children and children with autism were tested while they observed either an experimenter grasping a piece of food for eating or grasping a piece of paper for placing it into a container (Figure 15.5). The EMG activity of the mylohyoid muscle (MH), a muscle involved in opening of the mouth, was recorded. The results showed that in TD children, the observation of food grasping determined the activation of MH, while this activation was lacking in children with autism. In other words, while the observation of an action done by another individual intruded into the motor system of a TD observer, this intrusion was lacking in children with autism. This finding indicates that, in autism, the mirror mechanism is "dormant" during action observation and the immediate, experiential understanding of others' intention is absent.

Both autistic and TD children were also asked to perform the two actions described above (grasp to eat and grasp to place), while EMG of MH muscle was recorded. In TD children the muscle became active as soon they moved the arm to reach the food. In contrast, no MH muscle activation was observed during food reaching and grasping in autistic children. MH muscle activation appeared only when the children brought the food to their mouth. These data indicate on the one side that ASD children are unable to organize their motor acts into a unitary action characterized by a specific intention, on the other they show a deficit in the mirror mechanism reflected by the absence of motor activation of the muscles involved in the observed action.

These findings show an apparent contradiction between the cognitive capacities of the children to report the purpose of the experimenter's action and their lack of motor resonance with the action. In order to clarify this incongruence a further experiment was performed in which TD and autistic children were shown with goal-directed motor acts and asked to report what the actor was doing and why he was doing it (Boria, Fabbri-Destro, Cattaneo, Sparaci, Sinigaglia, Santelli, et al., 2009). These tasks test two different abilities; first to recognize a motor act (e.g. grasping an object) and second to understand the intention behind it (e.g. grasping to eat). The results showed that both TD and ASD children were able to recognize what the actor was doing, but ASD children failed in recognizing why the actor was doing it. ASD children systematically attributed to the actor the intention that could be derived by the semantics of the object, e.g. intention to cut when

scissors were shown, regardless of how the object was grasped. This finding indicates that ASD children interpret the behavior of others on the basis of the standard use of objects rather than of the behavior of the actors. ASD children therefore appear to have deficiencies in reading the intention of others from their behavior.

It may sound surprising that children with autism have no problem in recognizing the motor act of others (e.g. grasping) when this appears one of the major functions of the mirror mechanism. This, however, is easily explained if one take into consideration that this capacity could be solved by STS activity. The difference between STS and mirror mechanism in perception consists not in motor act recognition, but in the capacity of the mirror mechanism to generalize the observed motor act and to allow one to understand the intention behind it as if done by observer himself (Rizzolatti & Sinigaglia, 2010).

It is important to conclude on this point with a sentence by Marc Jeannerod (2004, p. 392): “Mere visual perception, without involvement of the motor system would only provide a description of the visible aspects of the movements of the agent, but it would not give precise information about the intrinsic components of the observed action which are critical for understanding what the action is about, what is its goal, and how to reproduce it”. This sentence beautifully describes how superficial and devoid of real understanding is the visual/inferential way of understanding others relative to that we defined as understanding “from the inside”.

Acknowledgements

This work was supported by ERC grant COGSYSTEMS to GR, by Regione-Università Emilia-Romagna and by Fondazione Banca Monte Parma.

References

- Altschuler, E. L., Vankov, A., Wang, V., Ramachandran, V. S., & Pineda JA. (1997). Person see, person do: Human cortical electrophysiological correlates of monkey see monkey do cells. Presented at a poster session, at the 27th Annual Meeting of the Society for Neuroscience, New Orleans.
- Augustine, J. R. (1996). Circuitry and functional aspects of the insular lobe in primates including humans. *Brain Research Reviews* 22: 229–44.
- Avikainen, S., Wohlschläger, A., Liuhanen, S., Hänninen, R., & Hari R. (2003). Impaired mirror-image imitation in Asperger and high-functioning autistic subjects. *Current Biology* 13: 339–41.
- Babiloni, C., Babiloni, F., Carducci, F., Cincotti, F., Coccozza, G., Del Percio, C., Moretti, D. V., & Rossini, P. M. (2002). Human cortical electroencephalography (EEG) rhythms during the observation of simple aimless movements: a high-resolution EEG study. *NeuroImage* 17: 559–72.
- Barracough, N. E., Xiao, D., Oram, M. W., & Perrett, D. I. (2006). The sensitivity of primate STS neurons to walking sequences and to the degree of articulation in static images. *Progress in Brain Research* 154: 135–48.
- Belmalih, A., Borra, E., Contini, M., Gerbella, M., Rozzi, S., & Luppino, G. (2007). A multiarchitectonic approach for the definition of functionally distinct areas and domains in the monkey frontal lobe. *Journal of Anatomy* 211: 199–211.
- Belmalih, A., Borra, E., Contini, M., Gerbella, M., Rozzi, S., & Luppino, G. (2009). Multimodal architectonic subdivision of the rostral part (area F5) of the macaque ventral premotor cortex. *Journal of Comparative Neurology* 512: 183–217.
- Bonini, L., Rozzi, S., Serventi, F. U., Simone, L., Ferrari, P. F., & Fogassi, L. (2010). Ventral premotor and inferior parietal cortices make contribution to action organization and intention understanding. *Cerebral Cortex* 20: 1372–85.
- Boria, S., Fabbri-Destro, M., Cattaneo, L., Sparaci, L., Sinigaglia, C., Santelli, E., Cossu, G., & Rizzolatti, G. (2009). Intention understanding in autism. *PLoS One* 4: e5596.

- Brass, M., Schmitt, R. M., Spengler, S. & Gergely, G. (2007). Investigating action understanding: inferential processes vs. action simulation. *Current Biology* 17: 2117–2121.
- Brodmann, K. (1909). Vergleichende Lokalisationslehre der Großhirnrinde.
- Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R. J., Zilles, K., Rizzolatti, G., & Freund, H. J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience* 13: 400–4.
- Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., Porro, C. A., & Rizzolatti, G. (2004a). Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study. *Journal of Cognitive Neuroscience* 16: 114–26.
- Buccino, G., Vogt, S., Ritzl, A., Fink, G. R., Zilles, K., Freund, H. J., & Rizzolatti, G. (2004b). Neural circuits underlying imitation learning of hand actions: An event-related fMRI study. *Neuron* 42: 323–34.
- Byrne, R. W. (2002). Seeing actions as hierarchically organized structures: Great ape manual skills. In A. N. Meltzoff & W. Prinz (Eds), *The Imitative Mind: Development, Evolution and Brain Bases* (pp. 122–40). Cambridge: Cambridge University Press.
- Caetano, G., Jousmäki, V., & Hari, R. (2007). Actor's and observer's primary motor cortices stabilize similarly after seen or heard motor actions. *Proceedings of the National Academy of Sciences* 104: 9058–62.
- Caggiano, V., Fogassi, L., Rizzolatti, G., Thier, P., & Casile, A. (2009). Mirror neurons differentially encode the peripersonal and extrapersonal space of monkeys. *Science* 324: 403–6.
- Caggiano, V., Fogassi, L., Rizzolatti, G., Pomper, J. K., Thier, P., Giese, M., & Casile, A. (2011). View-based encoding of actions in mirror neurons of area F5 in macaque premotor cortex. *Current Biology* 21: 144–8.
- Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience* 3: 1077–8.
- Carr, L., Iacoboni, M., Dubeau, M. C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences* 100: 5497–502.
- Carruthers, P., & Smith, P. K. (1996). *Theories of Theories of Mind*. Cambridge: Cambridge University Press.
- Caruana, F., Jezzini, A., Sbriscia-Fiochetti, B., Rizzolatti, G., & Gallese, V. (2011). Emotional and social behaviors elicited by electrical stimulation of the insula in the macaque monkey. *Current Biology* 21: 195–9.
- Caspers, S., Zilles, K., Laird, A. R., & Eickhoff, S. B. (2010). ALE meta-analysis of action observation and imitation in the human brain. *NeuroImage* 15: 1148–67.
- Cattaneo, L., Caruana, F., Jezzini, H., & Rizzolatti, G. (2009). Representation of goal and movements without overt motor behavior in the human motor cortex: a TMS study. *Journal of Neuroscience* 29: 11134–8.
- Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., & Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of the National Academy of Sciences* 104: 17825–30.
- Cochin, S., Barthelemy, C., Lejeune, B., Roux, S., & Martineau, J. (1998). Perception of motion and qEEG activity in human adults. *Electroencephalography and Clinical Neurophysiology* 107: 287–95.
- Cochin, S., Barthelemy, C., Roux, S., & Martineau, J. (1999). Observation and execution of movement: similarities demonstrated by quantified electroencephalography. *European Journal of Neuroscience* 11: 1839–42.
- Colby, C. L., Duhamel, J.-R., & Goldberg, M. E. (1993). Ventral intraparietal area of the macaque: anatomic location and visual response properties. *Journal of Neurophysiology* 69: 902–14.
- Crammond, D. J., & Kalaska, J. F. (2000). Prior information in motor and premotor cortex: activity during the delay period and effect on pre-movement activity. *Journal of Neurophysiology* 84: 986–1005.

- Damasio, A. R. (2003). Feelings of emotion and the self. *Annals of the New York Academy of Sciences* 1001: 253–61.
- Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., & Iacoboni, M. (2006). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* 9: 28–30.
- Darwin, C. (1872). *The Expression of Emotions in Man and Animals*. London: John Murray.
- de Lange, F. P., Spronk, M., Willems, R. M., Toni, I., & Bekkering, H. (2008). Complementary systems for understanding action intentions. *Current Biology* 18: 454–57.
- di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research* 91: 176–80.
- Dowell, L. R., Mahone, E. M., & Mostofsky, S. H. (2009). Associations of postural knowledge and basic motor skill with dyspraxia in autism: implication for abnormalities in distributed connectivity and motor learning. *Neuropsychology* 23: 563–70.
- Duhamel, J.-R., Colby, C. L., & Goldberg, M. E. (1998). Ventral intraparietal area of the macaque: congruent visual and somatic response properties. *Journal of Neurophysiology* 79: 126–36.
- Dziuk, M. A., Gidley Larson, J. C., Apostu, A., Mahone, E. M., Denckla, M. B., & Mostofsky, S. H. (2007). Dyspraxia in autism: association with motor, social, and communicative deficits. *Developmental Medicine & Child Neurology* 49: 734–9.
- Ekman, P. (1999). *Handbook of Cognition and Emotion*. Chichester: John Wiley & Sons, Ltd.
- Enticott, P. G., Kennedy, H. A., Rinehart, N. J., Tonge, B. J., Bradshaw, J. L., Taffe, J. R., Daskalakis, Z. J., & Fitzgerald, P. B. (2012). Mirror Neuron Activity Associated with Social Impairments but not Age in Autism Spectrum Disorder. *Biological Psychiatry* 71: 427–33.
- Fabbri-Destro, M., & Rizzolatti, G. (2008). Mirror neurons and mirror systems in monkeys and humans. *Physiology (Bethesda)* 23: 171–9.
- Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology* 73: 2608–11.
- Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science* 302: 662–7.
- Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M., & Rizzolatti, G. (1996). Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology* 76: 141–57.
- Filimon, F., Nelson, J. D., Hagler, D. J., & Sereno, M. I. (2007). Human cortical representations for reaching: mirror neurons for execution, observation, and imagery. *NeuroImage* 37: 1315–28.
- Frith, C. D., & Frith, U. (2007). Social cognition in humans. *Current Biology* 17: 724–32.
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119: 593–609.
- Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences* 8: 396–403.
- Gangitano, M., Mottaghy, F. M., & Pascual-Leone, A. (2001). Phase-specific modulation of cortical motor output during movement observation. *NeuroReport* 12: 1489–92.
- Gastaut, H., Terzian, H., & Gastaut, Y. (1952). Etude d'une activité électroencéphalographique méconnue: "Le rythme rolandique en arceau". *Mars Medical* 89: 296–310.
- Gazzola, V., Aziz-Zadeh, L., & Keysers, C. (2006). Empathy and the somatotopic auditory mirror system in humans. *Current Biology* 16: 1824–9.
- Gazzola V, & Keysers C. (2009). The observation and execution of actions share motor and somatosensory voxels in all tested subjects: single-subject analyses of unsmoothed fMRI data. *Cerebral Cortex* 19: 1239–55,
- Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage* 35: 1674–84.

- Gentilucci, M., Scandolara, C., Pigarev, I. N., & Rizzolatti, G. (1983). Visual responses in the postarcuate cortex (area 6) of the monkey that are independent of eye position. *Experimental Brain Research* 50: 464–8.
- Gerbella, M., Belmalih, A., Borra, E., Rozzi, S., & Luppino, G. (2010). Cortical connections of the macaque caudal ventrolateral prefrontal areas 45A and 45B. *Cerebral Cortex* 20: 141–68.
- Geyer, S., Matelli, M., Luppino, G., & Zilles, K. (2000). Functional neuroanatomy of the primate isocortical motor system. *Anatomy and Embryology* 202: 443–74.
- Ghaziuddin, M., & Butler, E. (1998). Clumsiness in autism and Asperger syndrome: a further report. *Journal of Intellectual Disability Research* 42: 43–8.
- Grafton, S. T., Arbib, M. A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations in humans by PET: 2. Observation compared with imagination. *Experimental Brain Research* 112: 103–11.
- Graziano, M. S., Yap, G. S., & Gross, G. (1994). Coding the visual space by premotor neurons. *Science* 266: 1054–7.
- Grèzes, J., Armony, J. L., Rowe, J., & Passingham, R. E. (2003). Activations related to “mirror” and “canonical” neurones in the human brain: an fMRI study. *NeuroImage* 18: 928–37.
- Gritti, A., Bove, D., Di Sarno, A. M., D’Addio, A. A., Chiapparò, S., & Bove, R. M. (2003). Stereotyped movements in a group of autistic children. *Functional Neurology* 18: 89–94.
- Hari, R. (2006). Action-perception connection and the cortical mu rhythm. *Progress in Brain Research* 159: 253–60.
- Heiser, M., Iacoboni, M., Maeda, F., Marcus, J., & Mazziotta, J. C. (2003). The essential role of Broca’s area in imitation. *European Journal of Neuroscience* 17: 1123–8.
- Hurley, S., & Charter, N. (2005). *Perspective on Imitation: From Neuroscience to Social Science*, Vol. 1. Cambridge: MIT Press.
- Hutchinson, W. D., Davis, K. D., Lozano, A. M., Tasker, R. R., & Dostrovsky, J. O. (1999). Pain-related neurons in the human cingulate cortex. *Nature Neuroscience* 2: 403–5.
- Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science* 286: 2526–8.
- Ishida, H., Nakajima, K., Inase, M., & Murata, A. (2009). Shared mapping of own and others’ bodies in visuotactile bimodal area of monkey parietal cortex. *Journal of Cognitive Neuroscience* 22: 83–96.
- Jastorff, J., Begliomini, C., Fabbri-Destro, M., Rizzolatti, G., & Orban, G. A. (2010). Coding observed motor acts: different organizational principles in the parietal and premotor cortex of humans. *Journal of Neurophysiology* 104: 128–40.
- Jeannerod, M. (1988). *The Neural and Behavioral Organization of Goal-directed Movements*. Oxford: Oxford University Press.
- Jeannerod, M. (2004). Action from within. *International Journal of Sport and Exercise Psychology* 2: 376–402.
- Kalaska, J. F., & Crammond, D. J. (1995). Deciding not to GO: neuronal correlates of response selection in a GO/NOGO task in primate premotor and parietal cortex. *Cerebral Cortex* 5: 410–28.
- Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child* 2: 217–50.
- Keller, G. B., & Hahnloser, R. H. (2009). Neural processing of auditory feedback during vocal practice in a songbird. *Nature* 457: 187–90.
- Kilner, J. M., Salenius, S., Baker, S. N., Jackson, A., Hari, R., & Lemon, R. N. (2003). Task-dependent modulations of cortical oscillatory activity in human subjects during a bimanual precision grip task. *NeuroImage* 18: 67–73.
- Kohen-Raz, R., Volkmar, F. R., & Cohen, D. J. (1992). Postural control in children with autism. *Journal of Autism and Developmental Disorders* 22: 419–32.
- Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science* 297: 846–8.

- Koski, L., Wohlschlager, A., Bekkering, H., Woods, R. P., Dubeau, M. C., Mazziotto, J. C., & Iacoboni, M. (2002). Modulation of motor and premotor activity during imitation of target-directed actions. *Cerebral Cortex* 12: 847–55.
- Kraskov, A., Dancause, N., Quallo, M. M., Shepherd, S., & Lemon, R. N. (2009). Corticospinal neurons in macaque ventral premotor cortex with mirror properties: a potential mechanism for action suppression? *Neuron* 64: 922–30.
- Krolak-Salmon, P., Henaff, M. A., Isnard, J., Tallon-Baudry, C., Guenot, M., Vighetto, A., Bertrand, O., & Mauguier, F. (2003). An attention modulated response to disgust in human ventral anterior insula. *Annals of Neurology* 53: 446–53.
- Kurth, F., Zilles, K., Fox, P., Laird, A., & Eickhoff, S. (2010). A link between the systems: functional differentiation and integration within the human insula revealed by meta-analysis. *Brain Structure and Function* 214: 519–34.
- Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J., & DeYoe, E. A. (2005). Distinct cortical pathways for processing tool vs. animal sounds. *Journal of Neuroscience* 25: 5148–58.
- Liepelt, R., Von Cramon, D. Y., & Brass, M. (2008). How do we infer other's goals from non stereotypic actions? The outcome of context-sensitive inferential processing in right inferior parietal and posterior temporal cortex. *NeuroImage* 43: 784–92.
- Lu, M. T., Preston, J. B., & Strick, P. L. (1994). Interconnections between the prefrontal cortex and the premotor areas in the frontal lobe. *Journal of Comparative Neurology* 341: 375–92.
- Luppino, G., Matelli, M., Camarda, R., & Rizzolatti, G. (1993). Corticocortical connections of area F3 (SMA-proper) and area F6 (pre-SMA) in the macaque monkey. *Journal of Comparative Neurology* 338: 114–40.
- Maeda, F., Kleiner-Fisman, G., & Pascual-Leone, A. (2002). Motor facilitation while observing hand actions: specificity of the effect and role of observer's orientation. *Journal of Neurophysiology* 87: 1329–335.
- Malle, B. F., Moses, J. L., & Baldwin, D. A. (2001). *Intentions and Intentionality: Foundations of Social Cognition*. Cambridge: MIT press.
- Martineau, J., Cochin, S., Magne, R., & Barthelemy, C. (2008). Impaired cortical activation in autistic children: is the mirror neuron system involved? *International Journal of Psychophysiology* 68: 35–40.
- Matelli, M., Luppino, G., & Rizzolatti, G. (1985). Patterns of cytochrome oxidase activity in the frontal agranular cortex of the macaque monkey. *Behavioral Brain Research* 18: 125–36.
- Matelli, M., Luppino, G., & Rizzolatti, G. (1991). Architecture of superior and mesial area 6 and the adjacent cingulate cortex in the macaque monkey. *Journal of Comparative Neurology* 311: 445–62.
- Merleau-Ponty M. (1962). *Phenomenology of Perception* (transl. C. Smith). London: Routledge.
- Mesulam, M. M., & Mufson, E. J. (1982). Insula of the Old World Monkey. III: Efferent cortical output and comments on function. *Journal of Comparative Neurology* 212: 38–52.
- Minshew, N. J., Sung, K., Jones, B. L., & Furman, J. M. (2004). Underdevelopment of the postural control system in autism. *Neurology* 63: 2056–61.
- Miyahara, M., Tsujii, M., Hori, M., Nakanishi, K., Kageyama, H., & Sugiyama T. (1997). Brief report: motor incoordination in children with Asperger syndrome and learning disabilities. *Journal of Autism and Developmental Disorders* 27: 595–603.
- Mooney, E. L., Gray, K. M., & Tonge, B. J. (2006). Early features of autism: Repetitive behaviors in young children. *European Child & Adolescent Psychiatry* 15: 12–18.
- Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology* 20: 750–6.
- Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V., & Rizzolatti, G. (1997). Object representation in the ventral premotor cortex (area F5) of the monkey. *Journal of Neurophysiology* 78: 2226–30.
- Muthukumaraswamy, S. D., & Johnson, B. W. (2004a). Primary motor cortex activation during action observation revealed by wavelet analysis of the EEG. *Clinical Neurophysiology* 115: 1760–6.

- Muthukumaraswamy, S. D., Johnson, B. W., & McNair, N. A. (2004b). Mu rhythm modulation during observation of an object-directed grasp. *Cognitive Brain Research* 19: 195–201.
- Nayate, A., Bradshaw, J. L., & Rinehart, N. J. (2005). Autism and Asperger's disorder: are they movement disorders involving the cerebellum and/or basal ganglia? *Brain Research Bulletin* 67: 327–34.
- Nelissen, K., Borra, E., Gerbella, M., Rozzi, S., Luppino, G., Vanduffel, W., Rizzolatti, G., & Orban, G. A. (2011). Action observation circuits in the macaque monkey cortex. *Journal of Neuroscience* 31: 3743–56.
- Nishitani, N., & Hari, R. (2000). Temporal dynamics of cortical representation for action. *Proceedings of the National Academy of Sciences* 97: 913–18.
- Noterdaeme, M., Mildenberger, K., Minow, F., & Amorosa, H. (2002). Evaluation of neuromotor deficits in children with autism and children with a specific speech and language disorder. *European Child & Adolescent Psychiatry* 11: 219–25.
- Oberman, L. M., Ramachandran, V. S., & Pineda, J. A. (2008). Modulation of mu suppression in children with autism spectrum disorders in response to familiar or unfamiliar stimuli: the mirror neuron hypothesis. *Neuropsychologia* 46: 1558–65.
- Peeters, R., Simone, L., Nelissen, K., Fabbri-Destro, M., Vanduffel, W., Rizzolatti, G., & Orban, G. A. (2009). The representation of tool use in humans and monkeys: common and uniquely human features. *Journal of Neurophysiology* 29: 11523–39.
- Penfield, W., & Faulk, M. E. (1955). The insula: further observations on its function. *Brain* 78: 445–70.
- Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., Chitty, A. J., Hietanen, J. K., & Ortega, J. E. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology* 146: 87–113.
- Perry, A., & Bentin, S. (2009). Mirror activity in the human brain while observing hand movements: a comparison between EEG desynchronization in the mu-range and previous fMRI results. *Brain Research* 1282: 126–32.
- Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., Williams, S. C., Bullmore, E. T., Brammer, M., & Gray, J. A. (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society B: Biological Sciences* 265: 1809–17.
- Prather, J. F., Peters, S., Nowicki, S., & Mooney, R. (2008). Precise auditory-vocal mirroring in neurons for learned vocal communication. *Nature* 451: 249–50.
- Press, C., Cook, J., Blakemore, S. J., & Kilner, J. (2011). Dynamic modulation of human motor activity when observing actions. *Journal of Neuroscience* 31: 2792–800.
- Prinz, W. (1987). Ideomotor action. In H. Heuer & A. Sanders (Eds), *Perspective on Perception and Action* (pp. 47–76). Hillsdale: Erlbaumpp.
- Prinz, W. (2002). Experimental approaches to imitation. In A. Meltzoff & W. Prinz (Eds), *The Imitative Mind: Development, Evolution, and Brain Bases* (pp. 143–62). Cambridge: Cambridge University Press.
- Qiu, A., Adler, M., Crocetti, D., Miller, M. I., & Mostofsky, S. H. (2010). Basal ganglia shapes predict social, communication, and motor dysfunctions in boys with autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry* 49: 539–51.
- Ramachandran, V. S. (2011). *The Tell-tale Brain. Unlocking the Mystery of Human Nature*. London: William Heinemann.
- Raos, V., Evangeliou, M. N., & Savaki, H. E. (2007). Mental simulation of action in the service of action perception. *Journal of Neuroscience* 27: 12675–83.
- Raos, V., Umiltà, M. A., Murata, A., Fogassi, L., & Gallese, V. (2006). Functional properties of grasping-related neurons in the ventral premotor area F5 of the macaque monkey. *Journal of Neurophysiology* 95: 709–29.
- Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Matelli M. (1988). Functional organization of inferior area 6 in the macaque monkey. II. Area F5 and the control of distal movements. *Experimental Brain Research* 71: 491–507.

- Rizzolatti, G., & Craighero, L. (2004). The mirror neuron system. *Annual Review of Neuroscience* 27: 169–92.
- Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996a). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research* 3: 131–41.
- Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. (1996b). Localization of grasp representation in humans by PET: 1. Observation vs. execution. *Experimental Brain Research* 111: 246–55.
- Rizzolatti, G., Ferrari, P. F., Rozzi, S., & Fogassi, L. (2006). The inferior parietal lobule: where action becomes perception. *Wiley: Novartis Foundation Symposium* 270: 129–40.
- Rizzolatti, G., & Luppino, G. (2001). The cortical motor system. *Neuron* 31: 889–901.
- Rizzolatti, G., Luppino, G., & Matelli, M. (1998). The organization of the cortical motor system: new concepts. *Electroencephalography and Clinical Neurophysiology* 106: 283–96.
- Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature Review Neuroscience* 11: 264–74.
- Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & van der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human Movement Science* 26: 525–54.
- Rozzi, S., Ferrari, P. F., Bonini, L., Rizzolatti, G., & Fogassi, L. (2008). Functional organization of inferior parietal lobule convexity in the macaque monkey: electrophysiological characterization of motor, sensory and mirror responses and their correlation with cytoarchitectonic areas. *European Journal of Neuroscience* 28: 1569–88.
- Royet, J. P., Plailly, J., Delon-Martin, C., Kareken, D. A., & Segebarth, C. (2003). fMRI of emotional responses to odors: Influence of hedonic valence and judgment, handedness, and gender. *NeuroImage* 20: 713–28.
- Sakreida, K., Schubotz, R. I., Wolfensteller, U., & von Cramon, D. Y. (2005). Motion class dependency in observers' motor areas revealed by functional magnetic resonance imaging. *Journal of Neuroscience* 25: 1335–42.
- Saygin, A. P., Wilson, S. M., Hagler, D. J., Bates, E., & Sereno, M. I. (2004). Point-light biological motion perception activates human premotor cortex. *Journal of Neuroscience* 24: 6181–8.
- Shepherd, S. V., Klein, J. T., Deaner, R. O., & Platt, M. L. (2009). Mirroring of attention by neurons in macaque parietal cortex. *Proceedings of the National Academy of Sciences* 106: 9489–94.
- Shmuelof, L., & Zohary, E. (2006). A mirror representation of others' actions in the human anterior parietal cortex. *Journal of Neuroscience* 26: 9736–42.
- 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: 1157–62.
- Sprengelmeyer, R., Rausch, M., Eysel, U. T., & Przuntek, H. (1998). Neural structures associated with recognition of facial expressions of basic emotions. *Proceedings of the Royal Society B: Biological Sciences* 265: 1927–31.
- Strafella, A. P., & Paus, T. (2000). Modulation of cortical excitability during action observation: a transcranial magnetic stimulation study. *NeuroReport* 11: 2289–92.
- Teitelbaum, P., Teitelbaum, O., Nye, J., Fryman, J., & Maurer, R. G. (1998). Movement analysis in infancy may be useful for early diagnosis of autism. *Proceedings of the National Academy of Sciences* 95: 13982–7.
- Théoret, H., Halligan, E., Kobayashi, M., Fregni, F., Tager-Flusberg, H., & Pascual-Leone, A. (2005). Impaired motor facilitation during action observation in individuals with autism spectrum disorder. *Current Biology* 15: R84–5.
- Tiihonen, J., Kajola, M., & Hari, R. (1989). Magnetic mu rhythm in man. *Neuroscience* 32: 793–800.
- Tomasello, M., & Call, J. (1997). *Primate Cognition*. Oxford: Oxford University Press.
- Ulloa, E. R., & Pineda, J. A. (2007). Recognition of point-light biological motion: mu rhythms and mirror neuron activity. *Behavioral Brain Research* 183: 188–94.

- Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). "I know what you are doing": A neurophysiological study. *Neuron* 32: 91–101.
- Umiltà, M. A., Escola, L., Intskirveli, I., Grammont, F., Rochat, M., Caruana, F., Jezzini, A., Gallese, V., & Rizzolatti, G. (2008). When pliers become fingers in the monkey motor system. *Proceedings of the National Academy of Sciences* 105: 2209–213.
- Vogt, S., Buccino, G., Wohlschlager, A. M., Canessa, N., Shah, N. J., Zilles, K., Eickhoff, S. B., Freund, H. J., Rizzolatti, G., & Fink, G. R. (2007). Prefrontal involvement in imitation learning of hand actions: Effects of practice and expertise. *NeuroImage* 37: 1371–83.
- Wheaton, L. A., Carpenter, M., Mizelle, J. C., & Forrester, L. (2008). Preparatory band specific premotor cortical activity differentiates upper and lower extremity movement. *Experimental Brain Research* 184: 121–6.
- Wicker, B., Keysers, C., Plailly, J., Royet, J-P, Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula. The common neural basis of seeing and feeling disgust. *Neuron* 3: 655–64.
- Williams, J. H. G., Whiten, A., Suddendorf, T., & Perrett, D. I. (2001). Imitation, mirror neurons and autism. *Neuroscience & Biobehavioral Reviews* 25: 287–95.
- Yuan, H., Perdoni, C., & He, B. (2010). Decoding speed of imagined hand movement from EEG. (2010a). Conference Proceedings of IEEE English Medical and Biological Society 2010, pp. 142–5.
- Zald, D. H., & Pardo, J. V. (2000). Functional neuroimaging of the olfactory system in humans. *International Journal of Psychophysiology* 36: 165–81.

Social neuropeptides in the human brain: Oxytocin and social behavior

Markus Heinrichs, Frances S. Chen,
and Gregor Domes

The neuropeptide oxytocin (OXT) plays a central role in mammalian social life. Its function in regulating social approach, social memory, and attachment has been documented in numerous species (Carter, 1998; Insel & Young, 2001; Winslow & Insel, 2004). In humans, oxytocin also regulates emotion reading, social stress-buffering, and trust (Heinrichs, von Dawans, & Domes, 2009; Meyer-Lindenberg, Domes, Kirsch, & Heinrichs, 2011). Recent studies on clinical populations suggest that impaired functioning of the oxytocin system may contribute to mental disorders associated with social deficits including autism, social anxiety disorder, borderline personality disorder, and schizophrenia (Meyer-Lindenberg et al., 2011). This chapter will review recent advances in human oxytocin research, focusing on how this “social neuropeptide” contributes to a core aspect of human sociality: the motivation and ability to understand the minds of others.

Neurophysiology of the oxytocin system

OXT is synthesized in magnocellular neurons in the paraventricular and supraoptic nuclei of the hypothalamus. It is processed along axonal projections to the posterior lobe of the pituitary, where it is stored in secretory vesicles and released into peripheral circulation. In addition to the release from axonal terminals, there is also dendritic release into the extracellular space, resulting in both local action and diffusion through the brain to reach distant targets (Ludwig & Leng, 2006). Furthermore, smaller OXT-producing parvocellular neurons in the paraventricular nucleus project directly to other regions in the brain including the amygdala, hippocampus, striatum, suprachiasmatic nucleus, bed nucleus of stria terminalis, and brainstem. In these regions, OXT acts as a neuromodulator or neurotransmitter. For example, OXT modulates neural populations in the central amygdala (Huber, Veinante, & Stoop, 2005; Viviani, Charlet, van den Burg, Robinet, Hurni, Abatis, et al., 2011). The distribution of OXT receptors in the brain, however, is not yet fully known. For an overview of studies on endogenous levels of OXT and human behavior, see Heinrichs et al. (2009). For an overview of the neurogenetic mechanisms of the OXT system, including neuroimaging studies, see Meyer-Lindenberg et al. (2011) and Kumsta and Heinrichs (2013).

Methods in human oxytocin research

Several methods are currently available for investigating the relationship between OXT and behavior. In combination with established behavioral and neuroimaging paradigms, these methods make it possible to document OXT's actions on the behavioral level as well as on the neural level. Each method's unique strengths and limitations influence its suitability for research on a specific target

population. The use of these methods in parallel has provided largely converging evidence on the relationship between OXT and behavior, although some important questions remain regarding whether data collected using the different methods can be interpreted in comparable ways.

Plasma and cerebrospinal fluid

Peripheral OXT levels can readily be measured in samples of blood plasma, and several studies suggest that plasma OXT levels correlate with psychological functioning. However, the relationship between peripheral neuropeptide levels and central nervous system availability of these neuropeptides is unclear (Anderson, 2006; Carter, Pournajafi-Nazarloo, Kramer, Ziegler, White-Traut, Bello, et al., 2007; Horvat-Gordon, Granger, Schwartz, Nelson, & Kivlighan, 2005; Landgraf & Neumann, 2004). Therefore, the interpretation of peripheral neuropeptide levels in psychological terms remains controversial (Heinrichs et al., 2009).

A second method for measuring OXT levels is through samples of cerebrospinal fluid (CSF). Relative to plasma, the OXT level in CSF is a more direct measure of the availability of the neuropeptide in the brain (Born, Lange, Kern, McGregor, Bickel, & Fehm, 2002) and is more directly relevant for behavioral effects and psychopathology than peripheral levels (Heinrichs & Domes, 2008). However, collecting CSF is a significantly more invasive procedure than collecting plasma, which limits the use of this method for research.

Molecular genetics

Molecular genetics allows for the study of naturally-occurring inter-individual variations in the OXT system along with their implications for human social behavior. Genetic material can be easily obtained through saliva samples, which increases the flexibility of this method and its suitability for potentially sensitive target populations such as infants. In humans, the oxytocin receptor is encoded by a gene located on chromosome 3p25 (Inoue, Kimura, Azuma, Inazawa, Takemura, Kikuchi, et al., 1994). Several recent studies have linked variation in the DNA sequence of the oxytocin receptor gene (*OXTR*) to social behavior and cognition. The two *OXTR* single nucleotide polymorphisms (SNPs) which have been most commonly linked to social behavior and cognition in recent studies are rs53576 (G/A) and rs2254298 (G/A) (Meyer-Lindenberg et al., 2011). As both SNPs are located in the intronic (non-coding) region of *OXTR*, further research on the molecular level will be necessary to clarify the precise relationship between these SNPs and the functionality of the oxytocin receptor (Kumsta & Heinrichs, 2013).

Intranasal administration

Intranasal administration of OXT provides the most direct method for studying the immediate effects of OXT on human brain activity and behavior. Generally, intranasal administration has been favored in recent research because it is the most effective and direct means available for experimentally manipulating neuropeptide levels in the human brain (Born et al., 2002; Heinrichs & Domes, 2008). Intranasal administration, unlike other methods, therefore allows causal inferences to be drawn between central OXT availability, and social behavior or cognition. Except where otherwise mentioned, the remainder of this chapter therefore focuses on studies conducted using this method.

Oxytocin and social approach

Social approach can be viewed as a psychological prerequisite for understanding the minds of others. OXT enhances the motivation to initiate and sustain social contact by dampening social

stress reactivity, increasing the motivation to engage in positive social interaction, and promoting emotions and cognitions associated with attachment and social bonding.

Overcoming social stress

Stressful social interactions, in particular with unfamiliar others, induce a well-characterized pattern of behavioral and physiological responses (often called the “fight-or-flight” response). The endocrine component of this response includes hypothalamus-pituitary-adrenal (HPA) axis activation and the secretion of CRH, ACTH, and cortisol. A number of studies suggest that OXT dampens this endocrine response during social stress. Breastfeeding women—in whom endogenous secretion of OXT is naturally increased—show muted cortisol responses to psychosocial stressors (Altemus, Deuster, Galliven, Carter, & Gold, 1995; Heinrichs, Meinlschmidt, Neumann, Wagner Kirschbaum Ehlert & Hellhammer, 2001). Healthy males randomly assigned to receive both social support and OXT during preparation for the “Trier Social Stress Test” (TSST) (Kirschbaum, Pirke, & Hellhammer, 1993) had lower cortisol responses, lower subjective anxiety, and higher reported calmness in the TSST than participants who received only social support, only OXT, or neither social support nor OXT (Heinrichs, Baumgartner, Kirschbaum, & Ehlert, 2003) (Figure 16.1). The physiological and psychological stress-buffering effects of OXT have been replicated and extended in other studies (de Oliveira, Zuairi, Graeff, Queiroz, & Crippa, 2011; Quirin, Kuhl, & Dusing, 2011).

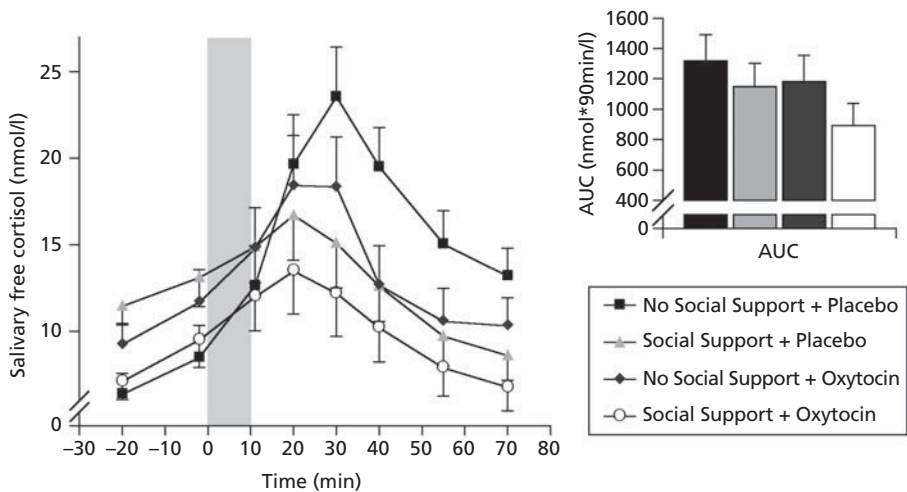


Figure 16.1 Mean salivary free cortisol concentrations (\pm SEM) during psychosocial stress exposure (TSST). Participants were randomly assigned to receive intranasal oxytocin (24 IU) or placebo and either no social support or social support from their best friend before stress. The shaded area indicates the period of the stress tasks (public speaking followed by mental arithmetic in front of a panel of evaluators). The areas under the individual response curves (AUC) represent cumulative cortisol release (calculated by aggregating data from 8 saliva sampling points) throughout the session. Significant interaction effects on cortisol were observed (social support by time effect, $p < .001$; social support by oxytocin by time effect, $P < 0.01$).

Adapted from Heinrichs, M., Baumgartner, T., Kirschbaum, C., & Ehlert, U. Biological Psychiatry, 54 (12), Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. pp. 1389–98. Copyright (2003), with permission from Elsevier.

Variation in the oxytocin receptor gene also seems to influence stress reactivity and responses to social support. Individuals homozygous for the G allele of rs53576 show a lower startle response than individuals with one or two copies of the A allele, although the stressor in this case was non-social (Rodrigues, Saslow, Garcia, John, & Keltner, 2009). The G allele of *OXTR* rs53576 has also been positively associated with the tendency to seek explicit social support during times of stress, at least in individuals for whom such behavior is culturally normative (Kim, Sherman, Sasaki, Xu, Chu, Ryu, et al., 2010). The same allele also interacts with social support to reduce physiological and psychological response to social stress in men (Chen, Kumsta, von Dawans, Monakhov, Ebstein, & Heinrichs, 2011c). Finally, receiving support and physical contact from a partner has been associated with an increase in plasma OXT (Grewen, Girdler, Amico, & Light, 2005), although on the other hand high levels of plasma OXT has also been associated with relationship stress in women (Taylor, Gonzaga, Klein, Hu, Greendale, & Seeman, 2006).

Overall, OXT seems to enhance the buffering effect of positive social interactions on stress responsiveness. The underlying biological mechanisms of this effect are now being investigated (Gamer & Buchel, 2011; Norman, Cacioppo, Morris, Malarkey, Berntson, & Devries, 2011). It is likely that the baseline sensitivity of the central nervous system (CNS) to OXT is influenced by significant events occurring early in life. Early parental separation stress, for example, has been shown to reduce men's sensitivity to intranasal OXT (Meinlschmidt & Heim, 2007).

On the neural level, the amygdala has been identified as one of the regions involved in the excitatory regulation of behavioral and endocrine stress responses (Dedovic et al., 2009). Increased activation of the amygdala has been consistently shown in response to aversive environmental stimuli (Pessoa & Adolphs, 2010, and in turn has been shown to contribute to the activation of the HPA axis in response to stress (Dedovic, Duchesne, Andrews, Engert, & Pruessner, 2009). The amygdala along with other subcortical regions contains a high density of OXT receptors and might thus represent a major target of central OXT (Gimpl & Fahrenholz, 2001). Furthermore, variation in the oxytocin receptor gene has been associated with amygdala volume (Furman, Chen, & Gotlib, 2011b; Inoue, Yamasue, Tochigi, et al., 2010) as well as amygdala activation during emotional face processing (Inoue et al., 2010). In animal studies, inhibitory oxytocinergic interneurons in the amygdala have been identified which suppress the output into autonomic target regions in the brain stem (Huber et al., 2005). Converging evidence from functional imaging studies has shown that intranasal OXT suppresses amygdala reactivity to aversive visual stimuli and reduces amygdala-brainstem coupling (Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008; Domes, Heinrichs, Glascher, Buchel, Braus, & Herpertz, 2007a; Gamer, Zurowski, & Buchel, 2010; Kirsch, Esslinger, Chen, Mier, Lis, Siddhanti, et al., 2005), which might be in part an underlying mechanism for the observed stress-reducing effect of OXT.

Trust and motivation to engage socially

In humans, one important indicator of psychological readiness for social approach is trust. In the first study to investigate the role of OXT in interpersonal trust (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005), participants' willingness to take social risks (in a trust game) vs. non-social risks (in a lottery game) was assessed. Participants who had received OXT showed greater trust relative to the placebo group (Figure 16.2). The effect was specific to the social context; OXT did not influence non-social risk-taking. The effect of OXT on trust has since been replicated in several follow-up studies (e.g. Mikolajczak, Gross, Lane, Corneille, de Timary, & Luminet, 2010a; Mikolajczak, Pinon, Lane, de Timary, & Luminet, 2010b).

A subsequent study suggested that OXT reduces negative behavioral responses to betrayal (Baumgartner et al., 2008). After playing several rounds of a trust game, participants were informed

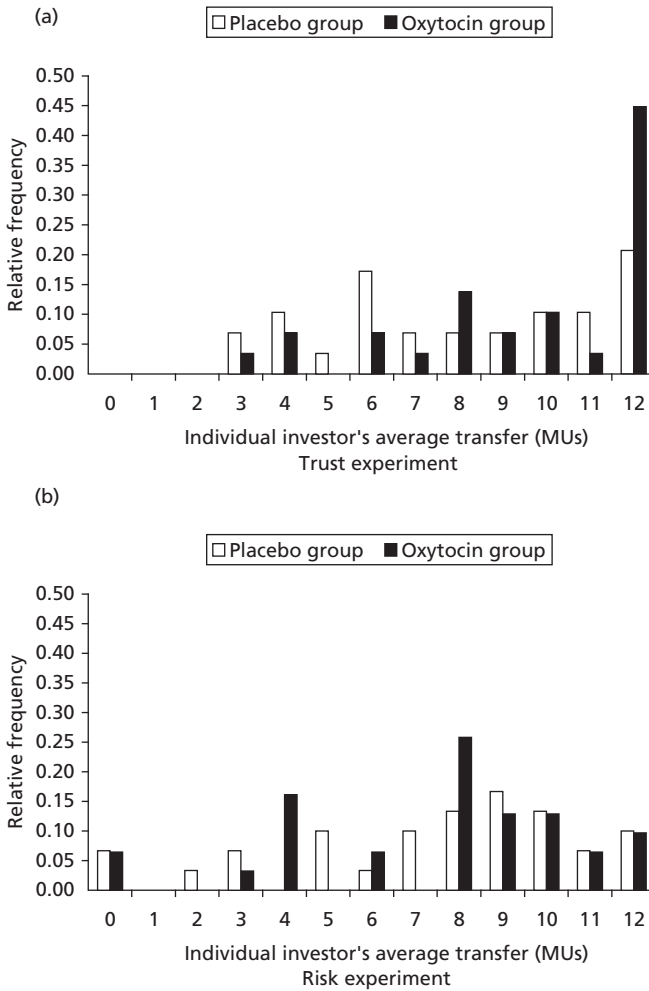


Figure 16.2 Transfers in the trust and risk experiments. Each observation represents the average transfer amount (in monetary units, MU) per investor across four transfer decisions. (a) Relative frequency of investors' average transfers in the oxytocin (filled bars) and placebo (open bars) groups in the trust experiment: subjects given oxytocin showed significantly higher transfer levels. (b) Relative frequency of investors' average transfers in the oxytocin (filled bars) and placebo (open bars) groups in the risk experiment: subjects in the oxytocin and the placebo group show statistically identical transfer levels.

Adapted from Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. Oxytocin increases trust in humans. *Nature* 435(7042) pp. 673–6. © 2011, with permission from Nature Publishing Group.

that their social partners had made selfish decisions that were disadvantageous to the participant (i.e. betrayed the participant's trust). Participants who had received intranasal OXT—in contrast to those who had received placebo—continued to make decisions indicative of sustained trust. OXT also increased the desire for future social interactions following an experience of inclusion in a game of “cyberball,” a virtual ball-tossing game designed to manipulate feelings of social inclusion

or exclusion, although it did not provide a buffer against the negative feelings associated with blunt social ostracism (Alvares, Hickie, & Guastella, 2010).

In addition to increasing trusting behavior, OXT has also been shown to increase generosity (Zak, Stanton, & Ahmadi, 2007) and to enhance perceived trustworthiness and attractiveness in facial expressions (Theodoridou, Rowe, Penton-Voak, & Rogers, 2009). High levels of plasma OXT also seem to correlate with trust and trustworthiness (Zak, Kurzban, & Matzner, 2005). Furthermore, it has been reported that OXT administration enhances envy and gloating in a social game (Shamay-Tsoory, Fischer, Dvash, Harari, Perach-Bloom, & Levkovitz, 2009), suggesting that OXT's effects on social behavior may not be modulated solely through prosocial emotions. The few studies that have directly investigated the specificity of these effects through the inclusion of both social and non-social stimuli suggest that the effects are more pronounced for social stimuli (Keri & Benedek, 2009; Norman, Cacioppo, Morris, Karelina, Malarkey, DeVries, et al., 2010). More recent research suggests that OXT effects are sensitive to the social context (Bartz, Zaki, Bolger, & Ochsner, 2011). Prior social information, such as brief prior face-to-face contact with a social partner, seems to enhance the effects of OXT on cooperative or prosocial behavior (Declerck, Boone, & Kiyonari, 2010). OXT effects on trust and cooperation also appear to be modulated by group membership, with stronger effects being observed among in-group members than out-group members (Chen, Kumsta, & Heinrichs, 2011b; De Dreu, Greer, Handgraaf, et al., 2010; De Dreu, Greer, Van Kleef, Shalvi, & Handgraaf, 2011).

Social bonding and attachment

In many mammalian species, OXT plays a crucial role in attachment and social bonding (Carter, 1998; Insel & Young, 2001; Young & Wang, 2004). In humans, OXT administration reduced plasma cortisol levels and increased positive communication in both men and women during a couple conflict (Ditzen, Schaer, Gabriel, Bodenmann, Ehlert, & Heinrichs, 2009), suggesting that central OXT facilitates human pair bonding in a manner parallel to that observed in prior animal studies. In men with an insecure attachment pattern, OXT also enhanced secure interpretations of ambiguous attachment-related scenarios (Buchheim, Heinrichs, George, Pokorny, Koops, Henningsen, et al., 2009). Genetic variability of the oxytocin receptor has been linked to attachment anxiety in adult females (Chen et al., 2011c). Secure attachment in adults is associated with healthy social functioning and reduced psychological responses to social stress (Ditzen, Schmidt, Strauss, Nater, Ehlert, & Heinrichs, 2008).

OXT has also been implicated in the human parent-infant attachment relationship. OXT administration increased fathers' responsiveness toward their toddlers during play (Naber, van Ijzendoorn, Deschamps, van Engeland, & Bakermans-Kranenburg, 2010). Levels of plasma OXT in pregnant women predict maternal behaviors and cognitions toward the infant after birth (Feldman, Weller, Zagoory-Sharon, & Levine, 2007). Variation in the oxytocin receptor gene has been linked to maternal sensitivity to child behavior (Bakermans-Kranenburg & van Ijzendoorn, 2008), as well as infant attachment security with the caregiver (Chen & Johnson, 2012).

The existing literature suggests that OXT promotes social approach behavior by reducing social stress reactivity, enhancing stress-buffering effects of positive social interaction, increasing motivation to engage in social interactions, and increasing trusting behavior, cooperation, and willingness to take social risks. OXT also promotes the maintenance of social relationships by enhancing secure attachment representations and increasing parents' responsiveness to their children. On the neural level, evidence suggests that OXT dampens amygdala reactivity to aversive social stimuli and reduces amygdala-brainstem coupling. OXT therefore promotes the initiation and maintenance of close social contact essential for learning about the mental states of others.

Oxytocin and social cognition

Reasoning about others' mental states depends critically on the ability to recognize and recall the emotional states experienced by others. OXT appears to regulate both emotion recognition (a cognitive component of empathy) and social memory.

Emotion recognition and empathy

OXT plays a central role in the recognition and processing of facial expressions of emotion. In the "Reading the Mind in the Eyes" Test (RMET), which was developed to assess the social cognitive abilities of adults with ASD (Baron-Cohen, Wheelwright, Hill Raste, & Plumb, 2001), participants are asked to judge other individuals' emotional or mental states based on photos of those individuals' eyes. Healthy men who had received OXT were more accurate on the RMET than those who had received placebo, particularly on the difficult items (Figure 16.3) (Domes, Heinrichs, Michel, Berger, & Herpertz, 2007b). The G allele of *OXTR* rs53576 has also been associated with better performance on the RMET (Rodrigues et al., 2009). The effects of OXT on emotion recognition accuracy have also been found to vary based on baseline individual differences in empathic abilities. In one study, intranasal OXT improved empathic accuracy (measured as the match between a participant's ratings of emotional states displayed by another individual in a film clip, and the displayed individual's own ratings his or her actual feelings) only in individuals with lower self-reported social-cognitive competency.

Studies which have investigated whether OXT selectively improves the recognition of specific emotions have thus far yielded mixed results. Some studies have documented that OXT specifically enhances processing of positive facial expressions (Di Simplicio, Massey-Chase, Cowen, & Harmer, 2009; Marsh, Yu, Pine, & Blair, 2010) or specifically decreases aversion to angry faces

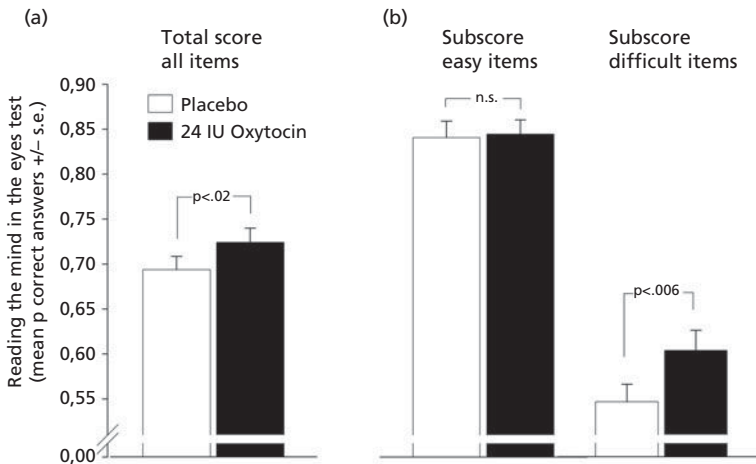


Figure 16.3 (a) Oxytocin improved performance in the RMET compared with placebo. (b) Performance in the RMET as a function of item difficulty: oxytocin improved performance on the difficult items and not on the easy items.

Adapted from Biological Psychiatry, 61 (6), Gregor Domes, Markus Heinrichs, Andre Michel, Christoph Berger and, Sabine C. Herpertz, Oxytocin Improves "Mind-Reading" in Humans, pp. 731–33, Copyright (2007), with permission from Elsevier.

(Evans, Shergill, & Averbeck, 2010). In contrast, other studies have reported improved recognition only of fearful faces after OXT administration (Fischer-Shofty, Shamay-Tsoory, Harari, & Levkovitz, 2010), or no effect on emotion recognition in a visual search task (Guastella, Carson, Dadds, Mitchell, & Cox, 2009a). Another recent study showed enhanced emotion recognition for both happy and angry faces at very short presentation durations of 17–83 milliseconds, suggesting that at least during the early stages of visual processing, OXT promotes recognition of a range of emotional expressions (Schulze, Lischke, Greif, Herpertz, Heinrichs, & Domes, 2011).

A related line of research has investigated the effects of OXT on visual attention to neutral and emotional faces, which is generally assumed to play a crucial role in the recognition of facial emotions (Adolphs, 2002). Three of these studies reported that OXT increased gazing time on the eye region compared with other parts of the face (Andari et al., 2010; Gamer et al., 2010; Guastella, Mitchell, & Dadds, 2008a), suggesting that improved facial emotion recognition after OXT treatment might be due at least in part to increased eye gaze. However, two other studies were not able to replicate increased eye gaze (Domes, Lischke, Berger, Grossmann, Hauenstein, Heinrichs, et al., 2010; Lischke, Berger, Prehn, Heinrichs, Herpertz, & Domes, 2011). It is possible that OXT interacts with baseline individual differences to enhance recognition of specific emotions and modulate eye gaze only under certain conditions.

Relative to studies investigating the effects of OXT on emotion recognition (generally theorized to represent a cognitive facet of empathy), research on OXT effects on emotional empathy, i.e. the vicarious feeling of an emotion, has so far been limited (Hurlemann, Patin, Onur, et al., 2010; Singer, Snozzi, Bird, et al., 2008). One recent study reported positive effects of intranasal OXT on emotional empathy but not cognitive empathy (Hurlemann et al., 2010), and another reported positive effects of OXT on “compassion-focused imagery” (Rockliff, Karl, McEwan, Gilbert, Matos, & Gilbert, 2011).

Social memory

Early work showed that OXT administration modulates semantic memory (Fehm-Wolfsdorf, Born, Voigt, & Fehm, 1984), and more recent studies suggest that OXT may more specifically modulate social memory. A study in healthy males showed that intranasal OXT selectively reduced implicit memory of socially relevant but not non-social words (Heinrichs, Meinlschmidt, Wippich, Ehlert, & Hellhammer, 2004). In another study, intranasal OXT selectively improved recognition memory for faces but not for nonsocial stimuli (Rimmele, Hediger, Heinrichs, & Klaver, 2009). Whether and how OXT administration influences memory for specific emotions remains unclear; existing studies have documented different effects depending on the timing of OXT administration. In one study, OXT administered after a learning task, improved memory after both a 30-minute and 1-day delay for faces that had displayed angry or neutral (and not happy) expressions during the learning task (Savaskan, Ehrhardt, Schulz, Walter, & Schachinger, 2008), although there was no effect of OXT on explicit memory for which of the specific facial expressions had been associated with specific identities. In another study, however, intranasal OXT administered before a learning task enhanced memory for happy faces over angry and neutral faces (Guastella, Mitchell, & Mathews, 2008b).

In summary, converging evidence suggests that OXT modulates the ability to decode and recall socially-relevant cues including those contained in facial expressions of emotion. These abilities are core components of a broader capacity to reason about others' mental states. Whether OXT has selective effects on recognition and memory of specific emotions remains unclear, as does the degree to which these effects are modulated by OXT-induced changes in patterns of visual attention. Targeted future research may help address these open questions.

Psychobiological therapy for social disorders

OXT's striking effects on human social behavior point to the potential therapeutic value of intranasal OXT for mental disorders characterized by social deficits. The possibility that atypical functioning of the OXT system contributes to specific mental disorders is supported by research comparing endogenous levels of OXT in healthy and clinical samples. Atypically low levels of plasma OXT have been observed in several mental disorders including autism spectrum disorders (ASD) (Green, Fein, Modahl, Feinstein, Waterhouse, & Morris, 2001), schizophrenia (Goldman, Marlow-O'Connor, Torres, & Carter, 2008; Keri, Kiss, & Kelemen, 2008), and depression (Cyranowski, Hofkens, Frank, Seltman, Cai, & Amico, 2008). In patients with depression, those with higher anxiety levels were found to have lower levels of plasma OXT (Scantamburlo, Hansenne, Fuchs, Pitchot, Maréchal, Pequeux, et al., 2007). Furthermore, variability of the oxytocin receptor gene has been linked to risk for autism (Jacob, Brune, Carter, Leventhal, Lord, & Cookjr, 2007; Lerer, Levi, Salomon, Darvasi, Yirmiya, & Ebstein, 2008; Wu, Jia, Ruan, Liu, Guo, Shuang, et al., 2005), although at least one null effect has also been reported (Tansey, Brookes, Hill, Cochrane, Gill, Skuse, et al., 2010). Because only a small fraction of intravenously-administered neuropeptide passes through the blood-brain barrier (Kang & Park, 2000), this method of OXT administration has limited applicability in clinical settings. Furthermore, intravenous infusion could potentially have side effects due to actions on hormone systems. Intranasal administration, which provides a direct pathway to the brain, currently shows the most promise as a clinical intervention methodology (Born et al., 2002; Heinrichs et al., 2009).

Several systematic, randomized control trials on the therapeutic effects of intranasal OXT treatment are now underway (see the ClinicalTrials website: <http://clinicaltrials.gov>). Although none of these trials are yet complete, preclinical studies in patients have already yielded promising initial results of a single dose of intranasal OXT on mental disorders characterized by social deficits. An overview of these studies is given below.

Autism spectrum disorder

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by speech and communication deficits, repetitive or compulsive behaviors in combination with restricted interests, and severe impairments in social functioning. In a recent study on adolescent males with ASD (Guastella, Einfeld, Gray, et al., 2010), a single dose of intranasal OXT improved performance on the RMET (Baron-Cohen et al., 2001). In another study (Andari, Duhamel, Zalla, Herbrecht, Leboyer, & Sirigu, 2010), intranasal OXT administered to adults with ASD increased social interactions and feelings of trust with fictitious partners in a simulated ball game ("cyberball"). In the same study, OXT administration also increased ASD patients' gazing time toward the eye region of facial photos. OXT administered intravenously has also been shown to improve understanding of emotional speech and decrease repetitive behaviors in individuals with ASD (Bartz & Hollander, 2008), although it should be noted that only a small fraction of intravenously administered OXT is thought to pass the blood-brain barrier. Overall, these studies suggest that OXT has therapeutic potential for core deficits associated with ASD, specifically by enhancing emotion recognition, reducing repetitive behaviors, and improving responsiveness to others.

Social anxiety disorder

Social anxiety disorder (SAD) is marked by extreme anxiety and discomfort in social settings and a fear of negative evaluation by others. After depression and alcoholism, it is the third most common mental health disorder in the United States (Kessler, McGonagle, Zhao, Nelson, Hughes,

Eshleman, et al., 1994). In one study, patients with SAD received either intranasal OXT or placebo a total of four times in combination with five weekly sessions of brief exposure intervention (Guastella, Howard, Dadds, Mitchell, & Carson, 2009b). Patients receiving OXT showed improved public speech performance, although—possibly due to the low frequency of sessions—a more generalized overall improvement in treatment outcome was not observed. In an fMRI study (Labuschagne, Phan, Wood, et al., 2010), SAD patients and healthy controls matched pictures of fearful, angry, and happy faces after OXT and placebo administration. In the placebo condition, patients with SAD exhibited amygdala hyperactivity to fearful faces relative to the control group. While OXT administration did not change amygdala reactivity to emotional faces in the control group, it dampened amygdala reactivity to fearful faces in the SAD group (Labuschagne et al., 2010). These findings suggest that OXT may have a specific effect on fear-related amygdala activity particularly when the amygdala is hyperactive as in SAD.

Borderline personality disorder

Borderline personality disorder (BPD) is characterized by emotional instability, impulsivity, identity diffusion, and dysfunctional social relationships. In BPD, the perception of rejection from a partner or close other often leads to angry outbursts or impulsive, suicidal, or self-injurious behavior. The suite of behaviors associated with BPD is theorized to be a result of disrupted functioning of the attachment and affiliative systems (Stanley & Siever, 2010). A recent pilot study suggests that OXT reduces stress reactivity in BPD patients during the TSST (Simeon, Bartz, Hamilton, et al., 2011). In another study, OXT administration decreased cooperative responses within the context of a social dilemma game; however, these effects were observed in a relatively small and mixed-sex sample of 14 adult BPD patients (Bartz, Simeon, Hamilton, et al., 2010a) and therefore warrant replication. Several clinical trials involving larger sample sizes are currently being conducted to examine the therapeutic value of OXT for patients with BPD (see clinicaltrials.gov).

Schizophrenia

Schizophrenia is a chronic brain disorder characterized by severely disorganized thought patterns, hallucinations and delusions, and disrupted affect. Schizophrenia has been associated with alterations in plasma OXT levels (Heinrichs et al., 2009). In animal models of schizophrenia, systematic OXT administration has been shown to have antipsychotic-like effects, including reversed pre-pulse inhibition deficits induced by amphetamine or the phencyclidine analogue MK 801 (Feifel & Reza, 1999). In humans, schizophrenia patients receiving a course of intranasal OXT for three weeks in addition to antipsychotics showed reduced positive and negative symptoms of schizophrenia (Feifel, Macdonald, Nguyen, et al., 2010). In another study (Goldman, Gomes, Carter, & Lee, 2011), emotion recognition in schizophrenia patients improved following administration of 20 IU of OXT in polydipsic relative to non-polydipsic patients, although performance fell in patients administered 10 IU of OXT.

Conclusions and future directions

This chapter has reviewed research linking OXT to both the motivation and ability to reason about the mental states of others. OXT promotes social approach by reducing social stress reactivity, increasing the motivation to interact with others, and enhancing trust and bonding. OXT also regulates social cognition, including the ability to recognize and recall others' emotional states. For individuals with pathologies in these domains, pharmacological intervention in the OXT system is a promising new angle for treatment. Initial studies with intranasal OXT in patients with mental

disorders characterized by social deficits (including autism, social anxiety disorder, borderline personality disorder, and schizophrenia) have been encouraging, especially given that these “social disorders” are notoriously difficult to treat or (as in the case of ASD) currently cannot be effectively treated at all.

Figure 16.4 depicts an integrative model of the relationships among the human central OXT system, social anxiety and stress, and social approach behavior. In this model, neuropeptide administration is seen as a means of supporting and enhancing psychotherapeutic interventions, rather than as an isolated alternative route to a cure. For example, treatment with intranasal OXT in combination with interaction-based psychotherapy may enhance patients’ willingness to interact socially (e.g. in cognitive-behavioral group therapy), as well as to confront feared social situations outside of therapy sessions. We propose the term “psychobiological therapy” for this novel integrative approach (Figure 16.4).

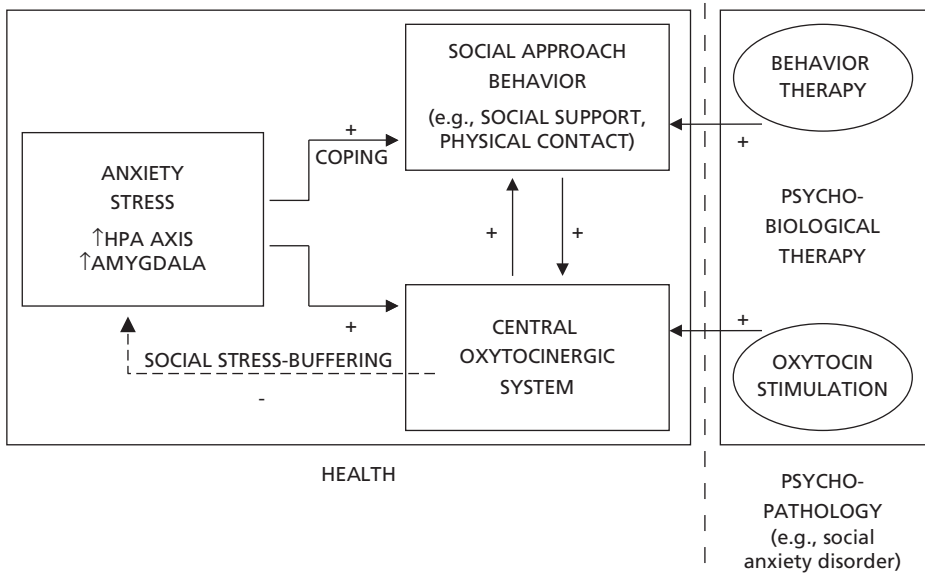


Figure 16.4 Integrative translational model of the interactions of the oxytocin system, social approach behavior, and social stress in humans. *Left side:* Social stress and social anxiety stimulate the amygdala–cingulate circuit and the hypothalamic–pituitary–adrenal (HPA) axis. In healthy individuals, stress and anxiety encourages social approach behavior as a coping strategy. It also stimulates oxytocin release, which further promotes social approach behavior. Furthermore, positive social interaction (e.g. physical contact) is itself associated with OXT release and therefore promotes continued social approach. OXT reduces amygdala and HPA axis reactivity to social stressors, and as such serves as an important mediator of the anxiolytic and stress-protective effects of positive social interaction (“social buffering”). *Right side:* Patients with mental and developmental disorders characterized by severe deficits in social interactions (e.g. autism, social anxiety disorder, borderline personality disorder) may benefit from novel “psychobiological therapy” approaches wherein psychotherapy is combined with administration of OXT or OXT receptor agonists.

Reprinted from *Progress in Brain Research*, 170, Markus Heinrichs and Gregor Domes, *Neuropeptides and social behaviour: effects of oxytocin and vasopressin in humans*, pp. 337–350, Copyright (2008), with permission from Elsevier.

Although significant progress has been made in understanding OXT's role in human social behavior, important details remain to be clarified in further research. A precise mapping of the distribution of OXT receptors in the human brain is crucial. This may be achieved through the development of specific radioactive labelling of neuropeptides in positron emission tomography, in combination with *in vitro* studies identifying OXT binding sites in the human brain (Loup, Tribollet, Dubois-Dauphin, & Dreifuss, 1991), as well as fMRI studies identifying brain areas responsive to OXT administration (Heinrichs & Domes, 2008; Meyer-Lindenberg, 2008). The mechanisms by which OXT, as well as OXT receptor agonists and antagonists, reach the brain following different forms of administration is an area for further study that may eventually lead to more effective methods for neuropeptide delivery. The development of non-peptidergic drugs (Decaux, Soupart, & Vassart, 2008) acting on OXT receptors is an important parallel goal. Targeted research on the relationship between central and peripheral release of oxytocin will be necessary to establish whether and how plasma OXT levels can be interpreted in terms of psychological function, in healthy subjects, as well as in patients. Further studies associating specific genetic variants with behavioral and neural responses to OXT administration may help clarify how naturally-occurring individual differences influence the functioning of the OXT system. Continuing research on oxytocin's role in human social behavior is contributing to a broader understanding of the neuroendocrinology of the social brain and may eventually lead to the development of more effective treatments for social disorders.

References

- Adolphs, R. (2002). Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behavioral Cognitive Neuroscience Reviews* 1: 21–62.
- Altemus, M., Deuster, P. A., Galliven, E., Carter, C. S., & Gold, P. W. (1995). Suppression of hypothalamic-pituitary-adrenal axis responses to stress in lactating women. *Journal of Clinical Endocrinology and Metabolism* 80: 2954–9.
- Alvares, G. A., Hickie, I. B., & Guastella, A. J. (2010). Acute effects of intranasal oxytocin on subjective and behavioral responses to social rejection. *Experimental and Clinical Psychopharmacology* 18: 316–21.
- Andari, E., Duhamel, J. R., Zalla, T., Herbrecht, E., Leboyer, M., & Sirigu, A. (2010). Promoting social behavior with oxytocin in high-functioning autism spectrum disorders. *Proceedings of the National Academy of Sciences of the United States of America* 107: 4389–94.
- Anderson, G. M. (2006). Report of altered urinary oxytocin and AVP excretion in neglected orphans should be reconsidered. *Journal of Autism and Developmental Disorders* 36: 829–30.
- Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2008). Oxytocin receptor (OXTR) and serotonin transporter (5-HTT) genes associated with observed parenting. *Social Cognitive and Affective Neuroscience* 3: 128–34.
- 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: 241–51.
- Bartz, J. A., & Hollander, E. (2008). Oxytocin and experimental therapeutics in autism spectrum disorders. *Progress in Brain Research* 170, 451–62.
- Bartz, J. A., Simeon, D., Hamilton, H., Kim, S., Crystal, S., Braun, A., et al. (2010a). Oxytocin can hinder trust and cooperation in borderline personality disorder. *Social Cognitive and Affective Neuroscience* 6: 556–63.
- Bartz, J. A., Zaki, J., Bolger, N., & Ochsner, K. N. (2011). Social effects of oxytocin in humans: context and person matter. *Trends in Cognitive Science* 15: 301–9.
- Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* 58: 639–50.

- Born, J., Lange, T., Kern, W., McGregor, G. P., Bickel, U., & Fehm, H. L. (2002). Sniffing neuropeptides: a transnasal approach to the human brain. *Nature Neuroscience* 5: 514–16.
- Buchheim, A., Heinrichs, M., George, C., Pokorny, D., Koops, E., Henningsen, P., O'Connor, M. F., & Gündel, H. (2009). Oxytocin enhances the experience of attachment security. *Psychoneuroendocrinology* 34: 1417–22.
- Carter, C. S. (1998). Neuroendocrine perspectives on social attachment and love. *Psychoneuroendocrinology* 23: 779–818.
- Carter, C. S., Pournajafi-Nazarloo, H., Kramer, K. M., Ziegler, T. E., White-Traut, R., Bello, D., & Schwartz, D. (2007). Oxytocin: behavioral associations and potential as a salivary biomarker. *Annals of the New York Academy of Sciences* 1098: 312–22.
- Chen, F. S., Barth, M. E., Johnson, S. L., Gotlib, I. H., & Johnson, S. C. (2011a). Oxytocin receptor (OXTR) polymorphisms and attachment in human infants. *Frontiers in Psychology* 2: 200.
- Chen, F. S., & Johnson, S. C. (2012). An oxytocin receptor gene variant predicts relationship anxiety in females and autism-spectrum traits in males. *Social Psychological and Personality Science* 3: 93–9.
- Chen, F. S., Kumsta, R., & Heinrichs, M. (2011b). Oxytocin and intergroup relations: Goodwill is not a fixed pie. *Proceedings of the National Academy of Sciences of the United States of America* 108: E45.
- Chen, F. S., Kumsta, R., von Dawans, B., Monakhov, M., Ebstein, R. P., & Heinrichs, M. (2011c). Common oxytocin receptor gene (OXTR) polymorphism and social support interact to reduce stress in humans. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 108: 19937–42.
- Cyranowski, J. M., Hofkens, T. L., Frank, E., Seltman, H., Cai, H. M., & Amico, J. A. (2008). Evidence of dysregulated peripheral oxytocin release among depressed women. *Psychosomatic Medicine* 70, 967–75.
- De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Handgraaf, M. J., Shalvi, S., Van Kleef, G., et al. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science* 328: 1408–1411.
- De Dreu, C. K., Greer, L. L., Van Kleef, G. A., Shalvi, S., & Handgraaf, M. J. (2011). Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences of the United States of America* 108: 1262–6.
- de Oliveira, D. C., Zuardi, A. W., Graeff, F. G., Queiroz, R. H., & Crippa, J. A. (2011). Anxiolytic-like effect of oxytocin in the simulated public speaking test. *Journal of Psychopharmacology* 26: 497–504.
- Decaux, G., Soupart, A., & Vassart, G. (2008). Non-peptide arginine-vasopressin antagonists: the vaptans. *Lancet* 371: 1624–32.
- Declerck, C. H., Boone, C., & Kiyonari, T. (2010). Oxytocin and cooperation under conditions of uncertainty: the modulating role of incentives and social information. *Hormones and Behavior* 57: 368–74.
- Dedovic, K., Duchesne, A., Andrews, J., Engert, V., & Pruessner, J. C. (2009). The brain and the stress axis: the neural correlates of cortisol regulation in response to stress. *NeuroImage* 47, 864–71.
- Di Simplicio, M., Massey-Chase, R., Cowen, P., & Harmer, C. (2009). Oxytocin enhances processing of positive vs. negative emotional information in healthy male volunteers. *Journal of Psychopharmacology* 23: 241–8.
- Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., & Heinrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological Psychiatry* 65: 728–31.
- Ditzen, B., Schmidt, S., Strauss, B., Nater, U. M., Ehlert, U., & Heinrichs, M. (2008). Adult attachment and social support interact to reduce psychological but not cortisol responses to stress. *Journal of Psychosomatic Research* 64: 479–86.
- Domes, G., Heinrichs, M., Glascher, J., Buchel, C., Braus, D. F., & Herpertz, S. C. (2007a). Oxytocin attenuates amygdala responses to emotional faces regardless of valence. *Biological Psychiatry* 62: 1187–90.
- Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007b). Oxytocin improves “mind-reading” in humans. *Biological Psychiatry* 61: 731–3.

- Domes, G., Lischke, A., Berger, C., Grossmann, A., Hauenstein, K., Heinrichs, M., & Herpertz, S. C. (2010). Effects of intranasal oxytocin on emotional face processing in women. *Psychoneuroendocrinology* 35: 83–93.
- Evans, S., Shergill, S. S., & Averbeck, B. B. (2010). Oxytocin decreases aversion to angry faces in an associative learning task. *Neuropsychopharmacology* 35: 2502–9.
- Fehm-Wolfsdorf, G., Born, J., Voigt, K. H., & Fehm, H. L. (1984). Human memory and neurohypophyseal hormones: opposite effects of vasopressin and oxytocin. *Psychoneuroendocrinology* 9: 285–92.
- Feifel, D., Macdonald, K., Nguyen, A., Cobb, P., Warlan, H., Galangue, B., et al. (2010). Adjunctive intranasal oxytocin reduces symptoms in schizophrenia patients. *Biological Psychiatry* 68: 678–80.
- Feifel, D., & Reza, T. (1999). Oxytocin modulates psychotomimetic-induced deficits in sensorimotor gating. *Psychopharmacology* 141: 93–8.
- Feldman, R., Weller, A., Zagoory-Sharon, O., & Levine, A. (2007). Evidence for a neuroendocrinological foundation of human affiliation: plasma oxytocin levels across pregnancy and the postpartum period predict mother-infant bonding. *Psychological Science* 18: 965–70.
- Fischer-Shofty, M., Shamay-Tsoory, S. G., Harari, H., & Levkovitz, Y. (2010). The effect of intranasal administration of oxytocin on fear recognition. *Neuropsychologia* 48: 179–84.
- Furman, D. J., Chen, M. C., & Gotlib, I. H. (2011). Variant in oxytocin receptor gene is associated with amygdala volume. *Psychoneuroendocrinology* 36: 891–7.
- Gamer, M., & Buchel, C. (2011). Oxytocin specifically enhances valence-dependent parasympathetic responses. *Psychoneuroendocrinology* 37: 87–93.
- Gamer, M., Zurowski, B., & Buchel, C. (2010). Different amygdala subregions mediate valence-related and attentional effects of oxytocin in humans. *Proceedings of the National Academy of Sciences of the United States of America* 107: 9400–5.
- Gimpl, G., & Fahrenholz, F. (2001). The oxytocin receptor system: Structure, function, and regulation. *Physiological Reviews* 81: 629–83.
- Goldman, M., Gomes, A. M., Carter, C. S., & Lee, R. (2011). Divergent effects of two different doses of intranasal oxytocin on facial affect discrimination in schizophrenic patients with and without polydipsia. *Psychopharmacology* 216: 101–10.
- Goldman, M., Marlow-O'Connor, M., Torres, I., & Carter, C. S. (2008). Diminished plasma oxytocin in schizophrenic patients with neuroendocrine dysfunction and emotional deficits. *Schizophrenia Research* 98: 247–55.
- Green, L., Fein, D., Modahl, C., Feinstein, C., Waterhouse, L., & Morris, M. (2001). Oxytocin and autistic disorder: alterations in peptide forms. *Biological Psychiatry* 50: 609–13.
- Grewen, K. M., Girdler, S. S., Amico, J., & Light, K. C. (2005). Effects of partner support on resting oxytocin, cortisol, norepinephrine, and blood pressure before and after warm partner contact. *Psychosomatic Medicine* 67: 531–8.
- Guastella, A. J., Carson, D. S., Dadds, M. R., Mitchell, P. B., & Cox, R. E. (2009a). Does oxytocin influence the early detection of angry and happy faces? *Psychoneuroendocrinology* 34: 220–5.
- Guastella, A. J., Einfeld, S. L., Gray, K. M., Rinehart, N. J., Tonge, B. J., Lambert, T. J., et al. (2010). Intranasal oxytocin improves emotion recognition for youth with autism spectrum disorders. *Biological Psychiatry* 67: 692–4.
- Guastella, A. J., Howard, A. L., Dadds, M. R., Mitchell, P., & Carson, D. S. (2009b). A randomized controlled trial of intranasal oxytocin as an adjunct to exposure therapy for social anxiety disorder. *Psychoneuroendocrinology* 34: 917–23.
- Guastella, A. J., Mitchell, P. B., & Dadds, M. R. (2008a). Oxytocin increases gaze to the eye region of human faces. *Biological Psychiatry* 63: 3–5.
- Guastella, A. J., Mitchell, P. B., & Mathews, F. (2008b). Oxytocin enhances the encoding of positive social memories in humans. *Biological Psychiatry* 64: 256–8.

- Heinrichs, M., Baumgartner, T., Kirschbaum, C., & Ehlert, U. (2003). Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. *Biological Psychiatry* 54: 1389–8.
- Heinrichs, M., & Domes, G. (2008). Neuropeptides and social behaviour: effects of oxytocin and vasopressin in humans. *Progress in Brain Research* 170: 337–50.
- Heinrichs, M., Meinlschmidt, G., Neumann, I., Wagner, S., Kirschbaum, C., Ehlert, U., & Hellhammer, D. H. (2001). Effects of suckling on hypothalamic-pituitary-adrenal axis responses to psychosocial stress in postpartum lactating women. *Journal of Clinical Endocrinology and Metabolism* 86: 4798–804.
- Heinrichs, M., Meinlschmidt, G., Wippich, W., Ehlert, U., & Hellhammer, D. H. (2004). Selective amnesic effects of oxytocin on human memory. *Physiology and Behavior* 83: 31–8.
- Heinrichs, M., von Dawans, B., & Domes, G. (2009). Oxytocin, vasopressin, and human social behavior. *Frontiers in Neuroendocrinology* 30: 548–57.
- Horvat-Gordon, M., Granger, D. A., Schwartz, E. B., Nelson, V. J., & Kivlighan, K. T. (2005). Oxytocin is not a valid biomarker when measured in saliva by immunoassay. *Physiology and Behavior* 84: 445–8.
- Huber, D., Veinante, P., & Stoop, R. (2005). Vasopressin and oxytocin excite distinct neuronal populations in the central amygdala. *Science* 308: 245–8.
- Hurlemann, R., Patin, A., Onur, O. A., Cohen, M. X., Baumgartner, T., Metzler, S., et al. (2010). Oxytocin enhances amygdala-dependent, socially reinforced learning and emotional empathy in humans. *Journal of Neuroscience* 30: 4999–5007.
- Inoue, H., Yamasue, H., Tochigi, M., Abe, O., Liu, X., Kawamura, Y., et al. (2010). Association between the oxytocin receptor gene and amygdalar volume in healthy adults. *Biological Psychiatry* 68: 1066–72.
- Inoue, T., Kimura, T., Azuma, C., Inazawa, J., Takemura, M., Kikuchi, T., Kubota, Y., Ogita, K., & Saji, F. (1994). Structural organization of the human oxytocin receptor gene. *Journal of Biological Chemistry* 269: 32451–6.
- Insel, T. R., & Young, L. J. (2001). The neurobiology of attachment. *Nature Reviews Neuroscience* 2: 129–36.
- Jacob, S., Brune, C., Carter, C., Leventhal, B., Lord, C., & Cookjr, E. (2007). Association of the oxytocin receptor gene (*OXTR*) in Caucasian children and adolescents with autism. *Neuroscience Letters* 417: 6–9.
- Kang, Y. S., & Park, J. H. (2000). Brain uptake and the analgesic effect of oxytocin—its usefulness as an analgesic agent. *Archives of Pharmaceutical Research* 23: 391–5.
- Keri, S., & Benedek, G. (2009). Oxytocin enhances the perception of biological motion in humans. *Cognitive, Affective & Behavioral Neuroscience* 9: 237–41.
- Keri, S., Kiss, I., & Kelemen, O. (2008). Sharing secrets: Oxytocin and trust in schizophrenia. *Social Neuroscience* 4: 1–7.
- Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H. U., & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychiatry* 51: 8–19.
- Kim H. S., Sherman, D. K., Sasaki, J. Y., Xu, J., Chu, T. Q., Ryu, C., Suh, E. M., Graham, K., & Taylor, S. E. (2010). Culture, distress, and oxytocin receptor polymorphism (*OXTR*) interact to influence emotional support seeking. *Proceedings of the National Academy of Sciences of the United States of America* 107: 15717–21.
- Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., Gruppe, H., Mattay, V. S., Gallhofer, B., & Meyer-Lindenberg, A. (2005). Oxytocin modulates neural circuitry for social cognition and fear in humans. *Journal of Neuroscience* 25: 11489–93.
- Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The “Trier Social Stress Test”—a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28: 76–81.
- Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature* 435: 673–6.
- Kumsta, R., & Heinrichs, M. (2013). Oxytocin, stress and social behavior: Neurogenetics of the human oxytocin system. *Current Opinion in Neurobiology*, 23, 11–16.

- Labuschagne, I., Phan, K. L., Wood, A., Angstadt, M., Chua, P., Heinrichs, M., et al. (2010). Oxytocin attenuates amygdala reactivity to fear in generalized social anxiety disorder. *Neuropsychopharmacology* 35: 2403–13.
- Landgraf, R., & Neumann, I. D. (2004). Vasopressin and oxytocin release within the brain: a dynamic concept of multiple and variable modes of neuropeptide communication. *Frontiers in Neuroendocrinology* 25: 150–76.
- Lerer, E., Levi, S., Salomon, S., Darvasi, A., Yirmiya, N., & Ebstein, R. P. (2008). Association between the oxytocin receptor (*OXTR*) gene and autism: relationship to Vineland Adaptive Behavior Scales and cognition. *Molecular Psychiatry* 13: 980–8.
- Lischke, A., Berger, C., Prehn, K., Heinrichs, M., Herpertz, S. C., & Domes, G. (2011). Intranasal oxytocin enhances emotion recognition from dynamic facial expressions and leaves eye-gaze unaffected. *Psychoneuroendocrinology* 37: 475–81.
- Loup, F., Tribollet, E., Dubois-Dauphin, M., & Dreifuss, J. J. (1991). Localization of high-affinity binding sites for oxytocin and vasopressin in the human brain. An autoradiographic study. *Brain Research* 555: 220–32.
- Ludwig, M., & Leng, G. (2006). Dendritic peptide release and peptide-dependent behaviours. *Nature Reviews Neuroscience* 7: 126–36.
- Marsh, A. A., Yu, H. H., Pine, D. S., & Blair, R. J. (2010). Oxytocin improves specific recognition of positive facial expressions. *Psychopharmacology* 209: 225–32.
- Meinlschmidt, G., & Heim, C. (2007). Sensitivity to intranasal oxytocin in adult men with early parental separation. *Biological Psychiatry* 61: 1109–11.
- Meyer-Lindenberg, A. (2008). Impact of prosocial neuropeptides on human brain function. *Progress in Brain Research* 170: 463–70.
- Meyer-Lindenberg, A., Domes, G., Kirsch, P., & Heinrichs, M. (2011). Oxytocin and vasopressin in the human brain: social neuropeptides for translational medicine. *Nature Reviews Neuroscience* 12: 524–38.
- Mikolajczak, M., Gross, J. J., Lane, A., Corneille, O., de Timary, P., & Luminet, O. (2010a). Oxytocin makes people trusting, not gullible. *Psychological Science* 21: 1072–4.
- Mikolajczak, M., Pinon, N., Lane, A., de Timary, P., & Luminet, O. (2010b). Oxytocin not only increases trust when money is at stake, but also when confidential information is in the balance. *Biological Psychology* 85: 182–4.
- Naber, F., van Ijzendoorn, M. H., Deschamps, P., van Engeland, H., & Bakermans-Kranenburg, M. J. (2010). Intranasal oxytocin increases fathers' observed responsiveness during play with their children: a double-blind within-subject experiment. *Psychoneuroendocrinology* 35, 1583–6.
- Norman, G. J., Cacioppo, J. T., Morris, J. S., Karelina, K., Malarkey, W. B., DeVries, A. C., & Berntson, G. G. (2010). Selective influences of oxytocin on the evaluative processing of social stimuli. *Journal of Psychopharmacology* 25: 1313–19.
- Norman, G. J., Cacioppo, J. T., Morris, J. S., Malarkey, W. B., Berntson, G. G., & Devries, A. C. (2011). Oxytocin increases autonomic cardiac control: moderation by loneliness. *Biological Psychology* 86: 174–80.
- Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: from a “low road” to “many roads” of evaluating biological significance. *Nature Reviews Neuroscience* 11: 773–83.
- Quirin, M., Kuhl, J., & Dusing, R. (2011). Oxytocin buffers cortisol responses to stress in individuals with impaired emotion regulation abilities. *Psychoneuroendocrinology* 36: 898–904.
- Rimmele, U., Hediger, K., Heinrichs, M., & Klaver, P. (2009). Oxytocin makes a face in memory familiar. *Journal of Neuroscience* 29: 38–42.
- Rockliff, H., Karl, A., McEwan, K., Gilbert, J., Matos, M., & Gilbert, P. (2011). Effects of intranasal oxytocin on “compassion focused imagery.” *Emotion* 11: 1388–96.

- Rodrigues, S., Saslow, L., Garcia, N., John, O., & Keltner, D. (2009). Oxytocin receptor genetic variation relates to empathy and stress reactivity in humans. *Proceedings of the National Academy of Sciences of the United States of America* 106: 21437–41.
- Savaskan, E., Ehrhardt, R., Schulz, A., Walter, M., & Schachinger, H. (2008). Post-learning intranasal oxytocin modulates human memory for facial identity. *Psychoneuroendocrinology* 33: 368–74.
- Scantamburlo, G., Hansenne, M., Fuchs, S., Pitchot, W., Maréchal, P., Pequeux, C., Ansseau, M., & Legros, J. J. (2007). Plasma oxytocin levels and anxiety in patients with major depression. *Psychoneuroendocrinology* 32, 407–10.
- Schulze, L., Lischke, A., Greif, J., Herpertz, S. C., Heinrichs, M., & Domes, G. (2011). Oxytocin increases recognition of masked emotional faces. *Psychoneuroendocrinology* 36: 1378–82.
- Shamay-Tsoory, S. G., Fischer, M., Dvash, J., Harari, H., Perach-Bloom, N., & Levkovitz, Y. (2009). Intranasal administration of oxytocin increases envy and schadenfreude (gloating). *Biological Psychiatry* 66: 864–70.
- Simeon, D., Bartz, J., Hamilton, H., Crystal, S., Braun, A., Ketay, S., et al. (2011). Oxytocin administration attenuates stress reactivity in borderline personality disorder: a pilot study. *Psychoneuroendocrinology* 36: 1418–21.
- Singer, T., Snozzi, R., Bird, G., Petrovic, P., Silani, G., Heinrichs, M. et al. (2008). Effects of oxytocin and prosocial behavior on brain responses to direct and vicariously experienced pain. *Emotion* 8: 781–91.
- Stanley, B., & Siever, L. J. (2010). The interpersonal dimension of borderline personality disorder: toward a neuropeptide model. *American Journal of Psychiatry* 167: 24–39.
- Tansey, K. E., Brookes, K. J., Hill, M. J., Cochrane, L. E., Gill, M., Skuse, D., Correia, C., Vicente, A., Kent, L., Gallagher, L., & Anney, R. J. (2010). Oxytocin receptor (*OXTR*) does not play a major role in the aetiology of autism: genetic and molecular studies. *Neuroscience Letters* 474: 163–7.
- Taylor, S. E., Gonzaga, G. C., Klein, L. C., Hu, P., Greendale, G. A., & Seeman, T. E. (2006). Relation of oxytocin to psychological stress responses and hypothalamic-pituitary-adrenocortical axis activity in older women. *Psychosomatic Medicine* 68: 238–45.
- Theodoridou, A., Rowe, A. C., Penton-Voak, I. S., & Rogers, P. J. (2009). Oxytocin and social perception: oxytocin increases perceived facial trustworthiness and attractiveness. *Hormones and Behavior* 56: 128–32.
- Viviani, D., Charlet, A., van den Burg, E., Robinet, C., Hurni, N., Abatis, M., Magara, F., & Stoop, R. (2011). Oxytocin selectively gates fear responses through distinct outputs from the central amygdala. *Science* 333: 104–7.
- Winslow, J. T., & Insel, T. R. (2004). Neuroendocrine basis of social recognition. *Current Opinion in Neurobiology* 14: 248–53.
- Wu, S., Jia, M., Ruan, Y., Liu, J., Guo, Y., Shuang, M., Gong, X., Zhang, Y., Yang, X., & Zhang, D. (2005). Positive association of the oxytocin receptor gene (*OXTR*) with autism in the Chinese Han population. *Biological Psychiatry* 58: 74–7.
- Young, L. J., & Wang, Z. (2004). The neurobiology of pair bonding. *Nature Neuroscience* 7: 1048–54.
- Zak, P. J., Kurzban, R., & Matzner, W. T. (2005). Oxytocin is associated with human trustworthiness. *Hormones and Behavior* 48: 522–7.
- Zak, P. J., Stanton, A. A., & Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PLoS One* 2: e1128.

Prenatal and postnatal testosterone effects on human social and emotional behavior

Bonnie Auyeung and Simon Baron-Cohen

Hormones are important chemical messengers that we use to regulate and control virtually all our physiological processes, from metabolism to activation of the immune system and the regulation of mood. They are also essential in the processes of reproduction, growth and development (Larsen, Kronenberg, Melmed, & Polonsky, 2002).

Animal studies of the effects of hormones have provided some of the clearest evidence for the role of various hormones in our bodies. More specifically, manipulation of glands producing particular hormones can have significant effects on physical development, as well as behavior (Christensen & Gorski, 1978; Collaer & Hines, 1995; Goy, Bercovitch, & McBair, 1988; Goy & McEwen, 1980; Hines, Davis, Coquelin, Goy, & Gorski, 1985). Mammals in particular have been widely studied, with castration (and thus reduction in gonadal hormones) being a common early experiment. These experiments show that hormones are essential to the sexual differentiation of both the body and the brain (see Collaer & Hines, 1995, for a review). It has long been recognized that castration of males at birth affects the development of masculine genitalia, while administration of androgens to females masculinizes their genitalia (Jost, 1970). Castrated males also usually show feminized neural development, cognition, and behavior; while females treated with androgen show masculinized neural development, cognition, and behavior. Similar experiments have been conducted in a wide range of mammals, comparing castrated males, normal males, normal females, and females treated with androgens on a range of sexually dimorphic features. These consistently demonstrate the importance of sex steroid hormones (testosterone in particular) in the development of the brain and behavior (Arnold & Gorski, 1984; Breedlove, 1994; Goy & McEwen, 1980; MacLusky & Naftolin, 1981; Williams & Meck, 1991).

While the effects of testosterone on non-human mammal sexual behavior have been extensively studied, there is now increasing evidence that this and other hormones also have a substantial effect on aspects of human social and emotional behavior (Baron-Cohen, Lutchmaya, & Knickmeyer, 2004; Cohen-Bendahan, van de Beek, & Berenbaum, 2005a; Hines, 2004). This chapter aims to review these findings by presenting a series of longitudinal studies designed to elucidate the behavioral effects of both prenatal and postnatal testosterone exposure in children and young adults. We discuss whether sex steroids, specifically testosterone, are also related to our social cognition, and specifically to understanding other minds.

Timing and critical periods

The timing of hormonal effects is crucial when studying lasting effects on development. There are generally thought to be two types of hormonal effects: organizational and activational (Phoenix, Goy, Gerall, & Young, 1959). Organizational effects are most likely to occur during early development when most neural structures are being established, producing permanent changes in the brain (Phoenix et al., 1959). In contrast, activational effects are short term and are dependent on current hormone levels. It is thought that organizational effects are maximal during certain critical periods of development. These are hypothetical time windows in which a tissue can be formed (Hines, 2004). Outside the critical period, the effect of the hormone will be limited, protecting the animal from disruptive influences. This means, for example, that circulating sex hormones necessary for adult sexual functioning do not cause unwanted alterations to tissues, despite the same hormones might having been essential in laying down cellular organization during the initial development of those tissues.

Animal research indicates that the critical period for sexual differentiation of the brain occurs when sex differences in serum testosterone are highest (Collaer & Hines, 1995). It is likely that this is an important period for sexual differentiation of the human brain as well. It is difficult to get accurate measurements of hormone levels for humans, but studies that have sampled fetal serum, plasma, and amniotic fluid during pregnancy have indicated that in typical human male fetuses, there is a surge in fetal testosterone (FT) levels between weeks 8–24 of gestation, peaking around week 16 (Abramovich & Rowe, 1973; Clements, Reyes, Winter, & Faiman, 1976; Reyes, Boroditsky, Winter, & Faiman, 1974; Reyes, Winter, & Faiman, 1973; Smail, Reyes, Winter, & Faiman, 1981). During this period, male fetuses produce more than 2.5 times the levels observed in females (Beck-Peccoz, Padmanabhan, Baggiani, Cortelazzi, Buscaglia, Medri, et al., 1991). From week 25 of gestation, there is then a decline, to barely detectable levels until birth. Therefore, we can assume that the significant effects of FT on development are likely to occur within this window. For typical human female fetuses, levels are generally very low throughout uterine development and during childhood (Hines, 2004; see Figure 17.1).

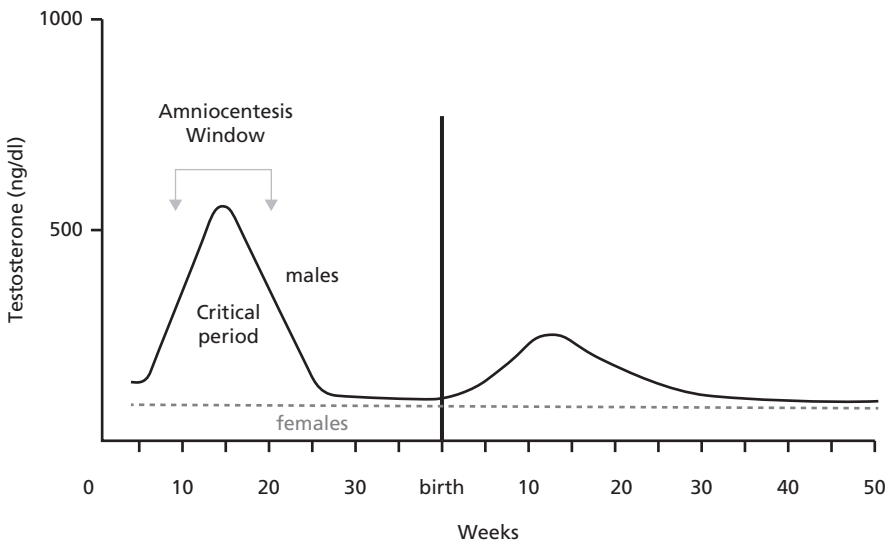


Figure 17.1 Sex differences in FT levels pre- and postnatally.

Prenatal hormone effects in humans

Studies in clinical conditions

Some naturally occurring medical conditions lead to atypical hormone levels. Since artificial manipulation of the hormone environment during critical periods of development is patently unethical, such medical conditions are a natural experiment for evaluating the impact of androgens and other hormones on development.

One such condition is congenital adrenal hyperplasia (CAH), a genetic mutation which causes excess adrenal production of androgen hormones (including testosterone and other hormones responsible for the development of masculinizing features) beginning prenatally in both males and females (New, 1998).

Behaviors showing large sex differences are the clearest candidates for studying effects of sex steroids (such as testosterone) on later development (Cohen-Bendahan et al., 2005a; Collaer & Hines, 1995; Hines, 2004). Studies of individuals with CAH have found that girls with the condition show masculinization of behavioral performance in activities such as spatial orientation, visualization, targeting, personality, cognitive abilities and sexuality (Hampson, Rovet, & Altmann, 1998; Hines, Fane, Pasterski, Matthews, Conway, & Brook, 2003; Resnick, Berenbaum, Gottesman, & Bouchard, 1986). While CAH provides an opportunity to investigate the effects of additional androgen exposure, the rare occurrence of CAH makes it difficult to obtain large enough groups and it is unclear if the research findings generalize to the wider population. Critics have also argued that CAH-related characteristics other than prenatal androgen exposure could be responsible for the atypical cognitive profiles observed in this population (Fausto-Sterling, 1992; Quadagno, Briscoe, & Quadagno, 1977).

Polycystic ovary syndrome (PCOS) is a common endocrine disorder in women, affecting one in 15 women. It is characterized by disruption of the ovulation cycle, a number of small cysts around the edge of their ovaries (polycystic ovaries), and excessive production and/or secretion of androgens (masculinizing hormones) referred to as “hyperandrogenism” (Norman, Dewailly, Legro, & Hickey, 2007). The daughters of women with PCOS show lower empathy quotient (EQ) scores, a measure on which typical girls usually show higher scores than boys (Palomba, Marotta, Di Cello, Russo, Falbo, Orio, et al., 2012). In the same study, daughters of women with PCOS also showed higher systemizing quotient (SQ) scores, a measure on which boys typically score higher than girls. These findings are consistent with the idea that PCOS increases androgen exposure *in utero*, and that this increased exposure leads the fetus to have more masculinized behavior in later life.

Studies using amniotic fluid measurements

Amniocentesis is the process of extracting a sample of amniotic fluid during the second trimester of pregnancy to detect clinical abnormalities in the fetus. Amniocentesis is typically performed during a relatively narrow time window, which coincides with the hypothesized critical period for human sexual differentiation (between approximately weeks 8 and 24 of gestation—see Figure 17.1; Hines, 2004). Samples taken in this way confirm that both male and female human fetuses produce testosterone, with male fetuses producing on average 2.5 times the levels observed in females.

Males produce testosterone from both the adrenal glands and the testes. The female fetus also produces androgens, but at much lower levels, mainly from the adrenal glands. In early prenatal life, this testosterone enters the amniotic fluid via diffusion through the fetal skin, and later enters the fluid via fetal urination (Robinson, Judd, Young, Jones, & Yen, 1977). Once produced, it is

carried by the blood to tissues around the body, and also crosses the blood–brain barrier to affect not just somatic tissue, but also neural development. Testosterone levels are also affected by other processes. For example, underproduction of aromatase may result in higher FT levels by impairing conversion of testosterone to estrogen (Abramovich, 1974). Similarly, dihydrotestosterone (DHT) is produced from testosterone and may be a stronger activator of the androgen receptor than testosterone itself (Larsen et al., 2002). Whilst these processes limit the conclusions we can draw from a snapshot measurement of FT level in the amniotic fluid, it is a useful starting point from which to develop our understanding.

A number of studies have linked elevated levels of FT in the amniotic fluid with the masculinization of certain behaviors, beginning shortly after birth. In particular, the Cambridge Child Development Project is an ongoing longitudinal study investigating the relationship between prenatal hormone levels and the development of later behavior (Baron-Cohen et al., 2004; Knickmeyer & Baron-Cohen, 2006). Mothers of participating children all underwent amniocentesis for clinical reasons. To date, these otherwise typically developing children have been tested postnatally at several time points. The findings from this study related to social and emotional understanding are discussed next, with particular reference to the child's development of a theory of mind.

Social development

Eye contact in infants is one of the earliest building blocks in social development and communication. The first study aimed to test if FT and estradiol levels showed any relationship to eye contact in a sample of 70 typically developing 12-month-old children (Lutchmaya, Baron-Cohen, & Raggatt, 2002a). Frequency and duration of eye contact were measured using videotaped sessions. Sex differences were found, with girls making significantly more eye contact than boys. The amount of eye contact inversely correlated with FT levels when the sexes were combined and also within the boys (Lutchmaya et al., 2002a). No relationships were observed between eye contact and estradiol levels. These results indicate that FT plays a role in shaping the neural mechanisms underlying social development (Lutchmaya et al., 2002a).

A second study tested if FT was related to the development of language, specifically examining the relationship between vocabulary size and FT and estradiol levels, as measured during amniocentesis. Vocabulary size was assessed at 18 and 24 months of age using the communicative development inventory, a self-administered checklist of words for parents to complete (Hamilton, Plunkett, & Shafer, 2000). Girls had significantly larger vocabularies than boys at both time points (Lutchmaya, Baron-Cohen, & Raggatt, 2002b) and results showed that levels of FT inversely predicted rate of vocabulary development (Lutchmaya et al., 2002b).

A follow-up of these children at 4 years of age used the children's communication checklist, a questionnaire designed to screen for communication difficulties in children 4–16 years of age (Bishop, 1998). The quality of social relationships subscale demonstrated an association between higher FT levels and poorer quality of social relationships for both sexes combined (but not within each sex). A lack of significant correlations within each sex may reflect the small sample size ($n = 58$).

Mentalizing/theory of mind

Mentalizing is the major focus of this volume and can be defined as the ability to put oneself into the mind of another person to infer what the person is thinking or feeling. It is also referred to as employing a theory of mind, or mindreading. The “moving geometric shapes” task was used: the children were asked to describe cartoons with two moving triangles whose interaction with each other suggested social relationships and psychological motivations (Knickmeyer, Baron-Cohen,

Raggatt, Taylor, & Hackett, 2006). Sex differences were observed, with girls using more mental and affective state terms to describe the cartoons compared with boys; however, no relationships between FT levels and frequency of mental or affective state terms were observed. Girls were found to use more intentional propositions than males, and a negative relationship between FT levels and frequency of intentional propositions was observed when the sexes were combined, as well as in boys alone. Boys used more neutral propositions than females. FT was a significant predictor of the frequency of neutral propositions when the sexes were combined.

Another method to measure mentalizing is the child version of the “Reading the Mind in the Eyes” test. This measure consists of 28 pictures from the eye region of the face, each depicting a mental state, some including subtle emotions (Baron-Cohen, Wheelwright, Spong, Scacchi, & Lawson, 2001). This revealed a significant, negative correlation with FT, with higher levels predicting lower mindreading capability. Within sex analyses revealed a significant negative correlation between FT and the Eyes test within both boys and girls (Chapman, Baron-Cohen, Auyeung, Knickmeyer, Taylor, & Hackett, 2006). The significance within each sex is important because it points to a more sensitive dependency on FT level than in the entire population, where boys have much higher levels of FT.

To understand the neural mechanisms by which this could take place we have investigated how FT affects brain development. The right temporo-parietal junction (RTPJ) is one region that is associated with tasks requiring one to think about other people’s thoughts and mental states (Saxe, 2010). Increases in FT predict increased gray matter in the RTPJ, and this brain region shows a male>female pattern of sexual dimorphism. This provides further clues suggesting a link between FT exposure and the neural development of mentalizing.

Gender-typical behavior

Children from the Cambridge Child Development Project were followed up using Pre-School Activities Inventory (PSAI). This is a standardized questionnaire measure of gender-typical play in both boys and girls. The PSAI includes 24 items and is completed by a parent to describe the child’s behavior. Higher scores reflect more male-typical behavior, and females with CAH obtain elevated (more male-typical) scores on the PSAI (Hines, Brook, & Conway, 2004), suggesting sensitivity to the effects of prenatal androgen exposure. A significant relationship exists between FT levels and sexually differentiated play behavior in both girls and boys (Auyeung, Baron-Cohen, Ashwin, Knickmeyer, Taylor, Hackett, et al., 2009a).

The Bem Sex Role Inventory (BSRI) is a questionnaire developed to measure feminine and masculine personality traits on the basis of cultural definitions of sex-typed social desirability (Bem, 1974). This is a 60-item (20 feminine, 20 masculine, and 20 non-gender related items) questionnaire. Examination of scores on this measure indicated that higher FT levels are associated with higher masculinity scores on the BSRI when boys and girls are examined together, and when girls are examined alone. No relationships are found between FT levels and scores on the femininity scale. Within sex results suggest that girls exposed to higher testosterone levels *in utero* are perceived as exhibiting more masculinized behavior (Auyeung, 2008).

Empathy

Empathy is the drive to identify another person’s emotions and thoughts, and to respond to these with an appropriate emotion (Baron-Cohen, 1999). This is an aspect of social interaction where females show a strong advantage. Sex differences in the precursors of empathy are seen from birth, with female babies showing a stronger preference for looking at social stimuli (faces) 24 hours after

birth (Connellan, Baron-Cohen, Wheelwright, Batki, & Ahluwalia, 2000), and more eye contact at 12 months of age (Lutchmaya et al., 2002a). Girls also tend to show more comforting, sad expressions or sympathetic vocalizations than boys when witnessing another's distress as early as 1 year of age (Hoffman, 1977).

Girls generally scored higher than boys on the Empathy Quotient-Child Version (EQ-C) at ages 6–8 years, which is by parent-report. A significant negative correlation between FT levels and EQ-C score is observed when the sexes are combined, and also within boys alone (Chapman et al., 2006).

Autistic traits

Autism Spectrum Conditions (ASC) are a group of related conditions characterized by impairments in reciprocal social interaction and communication, alongside strongly repetitive behaviors and unusually narrow interests (APA, 1994). Autism is much more prevalent in males than females (Chakrabarti & Fombonne, 2005; Gillberg, Cederlund, Lamberg, & Zeijlon, 2006), so the possibility that androgens may have a role to play in the etiology of these conditions has been explored.

Studies have examined the effects of FT on the later development of autism and autistic traits. In the first of these studies, autistic traits were measured using the Quantitative Checklist for Autism in Toddlers (Q-CHAT) (Allison, Baron-Cohen, Wheelwright, Charman, Richler, Pasco, et al., 2008). The Q-CHAT questionnaire was completed by mothers who had also undergone amniocentesis, providing measurements of FT level and fetal estradiol (FE)—a second hormone which forms prenatally from testosterone and is considered to be the most biologically active estrogen (Collaer & Hines, 1995). Samples of postnatal testosterone (PT) levels were also taken from saliva at 3–4 months of age in a small sample of these children. The study revealed a significant sex difference in autistic traits, with boys scoring higher (indicating more autistic traits) than girls. Q-CHAT scores were predicted by FT levels only, with both sex and the FT/Sex interaction excluded from the model (Auyeung, Taylor, Hackett, & Baron-Cohen, 2010).

The relationship between FT and Q-CHAT score was also visible within in the subset of children who participated in the follow up study measuring postnatal testosterone (PT) levels at 3–4 months. However, no relationships between FE, PT levels and Q-CHAT scores were observed. In addition, FE and PT levels showed no sex differences or relationships with FT levels (Auyeung, et al., 2013).

FT measurements were also directly evaluated against a child's score on the Childhood Autism Spectrum Test (CAST) (Scott, Baron-Cohen, Bolton, & Brayne, 2002; Williams, Scott, Stott, Allison, Bolton, Baron-Cohen, et al., 2005) and the Autism Spectrum Quotient-Child Version (AQ-Child) (Auyeung, Baron-Cohen, Wheelwright, & Allison, 2008). The CAST is a validated and widely used autism screening measure used to detect who is at risk for ASC. The AQ-Child is a measure that quantifies autistic traits and has been used widely in research.

FT levels are positively associated with higher scores (indicating greater number of autistic traits) on both the CAST and the AQ-Child. For the AQ-Child, this relationship is seen within both males and females as well as when the sexes are combined, suggesting this is an effect of FT, rather than an effect of sex. The relationship between CAST scores and FT is also seen within boys, but not girls (Auyeung, Baron-Cohen, Chapman, Knickmeyer, Taylor, & Hackett, 2009b).

Summary of the Cambridge Child Development Project

Table 17.1 describes the measures used in the Cambridge Child Development Project to identify sex differences in behavior and the links with FT for boys and girls together. For each measure, the

Table 17.1 Cambridge Child Development Project

Characteristic	Measure	Child's age	Sex difference	Results
Eye contact	Frequency	12 months	Yes (F>M)	Higher FT predicts less eye contact
Vocabulary size	Communicative development inventory	18–24 months	Yes (F>M)	Higher FT predicts smaller vocabulary size
Social relationships	Children's communication checklist	4 years	Yes (F>M)	Higher FT predicts poorer quality of social relationships
Mental and affective language	Intentional propositions	4 years	Yes (F>M)	Higher FT predicts less mental and affective language
Mindreading	Reading the mind in the eyes	6–9 years	No	Higher FT predicts poorer mindreading
Gender-typical play	PSAI	6–9 years	Yes (F<M)	Higher FT predicts more male-typical play preferences
Gender-role behavior	BSRI	6–9 years	Yes (F<M)	Higher FT predicts more male-typical sex role characteristics
Empathy	Empathy quotient	6–9 years	Yes (F>M)	Higher FT is associated with less empathy
Autistic traits	Q-CHAT	18–24 months	Yes (F<M)	Higher FT predicts more autistic traits
Autistic traits	AQ-Child	6–10 years	Yes (F<M)	Higher FT predicts more autistic traits
Autistic traits	CAST	6–10 years	Yes (F<M)	Higher FT predicts more autistic traits

direction of the sex differences (if present) is shown. The final column indicates whether FT levels (independent of sex) were a significant predictor in the associated regression analyses.

Limitations of measuring prenatal exposure to hormones in amniotic fluid

The findings presented in Table 17.1 make use of testosterone levels in amniotic fluid (sampled via amniocentesis). The benefit of this method is that it provides a sample that is close to the fetus and is collected as part of normal clinical practice for mothers thought to be at risk of complications during pregnancy or birth. Amniocentesis is generally also conducted in a fairly narrow time window, aiding repeatability of measurements. Ideally, it would be most useful to make direct measurements of testosterone at regular intervals throughout gestation and into postnatal life. Even for amniocentesis, it is not currently possible to obtain repeated samples of FT because the procedure carries a risk of causing miscarriage (about 1%) (d’Ercole et al., 2003; Sangalli et al., 2004). It is also known that hormones fluctuate during the day and between days, even in fetuses (Seron-Ferre et al., 1993; Walsh et al., 1984).

Given the estimated timeline for testosterone secretion, the most promising time to measure FT is probably at prenatal weeks 8–24 (Smail, Reyes, Winter, & Faiman, 1981), but this is still a relatively wide range. Research in non-human primates has also shown that androgens masculinize different behaviors at different times during gestation, suggesting different behaviors may also have different sensitive periods for development (Goy et al., 1988).

For all these reasons, the inferences we can therefore draw about the single measurement of FT are necessarily limited. At the same time, a significant correlation between amniotic FT and a behavior should represent a conservative estimate of the potential effect of FT exposure on that behavior.

Human behavior is complex, and biological, social or cultural factors are continuously interacting, making it challenging to investigate the causes of behavior. To the extent that social factors have been controlled within the experiments presented above, these were restricted to demographic variables such as maternal age, parental education, and number of siblings, and behaviors and traits are likely to be influenced by a range of social factors that have not been measured in these studies.

Postnatal hormone effects in humans

Studies of current (activational) hormones

Studies of postnatal hormone exposure have examined the effects of current (or activational) hormones. One obvious example of hormonal variation is the menstrual cycle, but the other obvious example of postnatal hormone exposure is during puberty. Studies in non-human mammals have investigated whether changes during puberty represent a critical period for the effects of steroid hormones. Gonadectomy in male ferrets before puberty, but after the early critical period does not affect sexual development when these animals are treated with testosterone in adulthood (Baum & Erskine, 1984). Early steroid hormone deprivation results in systemic reduction in sensitivity to later androgen effects (Gotz & Dorner, 1976). More recent findings suggest that steroid hormones during puberty have an activational effect on brain development (Schulz, Molenda-Figueira, & Sisk, 2009). These results indicate that although the critical window during perinatal development is vitally important for early sexual differentiation of the brain, the pubertal period also plays a large role in “fine-tuning” the organizational effects of steroid hormones (Romeo, Richardson, & Sisk, 2002).

During puberty, changes occur in adrenal androgens, rapid growth in body size, fat composition, and the development of secondary sex characteristics (Forbes & Dahl, 2010). Studies examining the relationships between puberty, hormone changes and the effects on social cognition and emotion have been relatively few. This is because onset of puberty varies greatly between individuals as well as between sexes, so recruitment of appropriate age groups can be difficult. In addition, there is little research in this age group due to the discomfort and embarrassment associated with trying to obtain reliable and accurate information on sexual maturation. Other studies have relied on parental or self-report measures of puberty which can have difficulties (Petersen, Crockett, Richards, & Boxer, 1988). It is also hard to disentangle the physical aspects of maturation from the co-occurring social changes associated with this age group.

Much of what is known about adolescent development in humans comes from studies that do not specifically include biological measures of pubertal development (such as hormone levels). For example, brain regions such as the rostral prefrontal cortex which are involved in executive functions are still developing during adolescence (Dumontheil, Burgess, & Blakemore, 2008). A study that used a narrow age range and measures of pubertal development showed a positive correlation

between pubertal development and an increased tendency toward sensation-seeking (Martin, Kelly, Rayens, Brogli, Brenzel, Smith, et al., 2002). The increase in sensation-seeking during puberty may relate to the increase in risk-taking observed in adolescents, which seems to decline in adulthood (Zuckerman, 1971). How the development of these systems is related to the effect of puberty or changes in steroid hormones is not known.

Efforts have been made to examine the links between prenatal and activational hormone effects on behavior in same-sex and opposite-sex twins, the assumption being that girls from pairs of opposite-sex twins are exposed to higher levels of prenatal testosterone, compared with same-sex twin girls (Cohen-Bendahan, Buitelaar, van Goozen, Orlebeke, & Cohen-Kettenis, 2005b). Such studies control for postnatal environmental effects by comparing data with similar measurements of same-sex female twins. The activational effects of testosterone are assessed using salivary testosterone measures in addition to a measure of pubertal status using the Tanner drawings (Tanner, 1962). Although there is some evidence of associations between free testosterone levels and personality traits (such as aggressive impulses and boredom susceptibility in boys, and experience seeking and extraversion in girls), no clear associations between circulating testosterone levels and behavioral traits are apparent.

More recently, sex differences have been observed in the relationship between circulating testosterone levels using bloodspot samples and thickness in areas of the brain associated with high androgen receptor density (including the left inferior parietal lobule, middle temporal gyrus, calcarine sulcus and right lingual gyrus; Bramen, Hranilovich, Dahl, Chen, Rosso, Forbes, et al., 2012). These findings provide new evidence for the role of testosterone in pubertal structural brain development and sexual differentiation. However, further work is needed to ascertain how these changes may relate to social, cognitive and emotional development.

Studies of testosterone administration

The majority of findings discussed so far have relied on observations in clinical conditions characterized by atypical exposure to hormones or by obtaining samples of amniotic fluid, blood, or saliva to measure hormone levels and relating these to measurements of interest. In some cases, it is also possible to study the effects of directly altering circulating hormone levels (though prenatal manipulation of hormone levels would be unethical). Recent studies in adult women have used a sublingual administration of testosterone, leading to a short-term large increase in circulating testosterone. Using this method, a series of studies have examined the effects of a single dose of testosterone vs. placebo on social and emotional behavior (see Bos, Panksepp, Bluthe, & van Honk, 2012b, for a review).

Administration studies have shown that testosterone decreases theory of mind and facial emotion recognition in these women. Using the "Reading the Mind in the Eyes" test, a measure examining subtle emotion and mental states from pictures of the eye region, testosterone administration led to lower scores compared with placebo (van Honk, Schutter, Bos, Kruijt, Lentjes, & Baron-Cohen, 2011). Interestingly, the 2D:4D digit length ratios (thought to be a proxy for prenatal hormone exposure) of the women tested in this study predicted approximately 50% of the variance in the effect of testosterone on task performance. The authors suggest that the testosterone administration effect may be primed by prenatal exposure to testosterone (van Honk et al., 2011).

Testosterone administration has also been shown to decrease recognition of angry expressions, and the authors hypothesize that testosterone may reduce the recognition of social threat, which may point toward a role for testosterone in social aggression (van Honk & Schutter, 2007). Angry faces may be an implicit signal of threat or competition, and testosterone administration has also been shown to increase gaze to the eye region of threatening faces that are viewed unconsciously, suggesting a role for testosterone in implicit social-dominance (Terburg, Aarts, & van Honk, 2012).

Testosterone has also been shown to reduce empathic facial imitation (Hermans, Putman, & van Honk, 2006). In an fMRI study, testosterone administration activated areas such as the orbitofrontal cortex and amygdala (both considered to be emotion processing regions) when looking at angry vs. happy facial expressions, again suggesting a role for testosterone in social threat (Hermans, Ramsey, & van Honk, 2008). A recent fMRI study also suggests that administration of testosterone alters functional connectivity between brain regions when looking at social stimuli. Testosterone (vs. placebo) decreases connectivity between the amygdala and orbitofrontal cortex (OFC) (Bos, Hermans, Ramsey, & van Honk, 2012a), and amygdala activation shifts away from the OFC, towards the thalamus (van Wingen, Mattern, Verkes, Buitelaar, & Fernandez, 2010).

Testosterone is also related to trust. Administration of testosterone is related to rating pictures as being less trustworthy compared with placebo, even when baseline testosterone levels do not differ (Bos, Terburg, & van Honk, 2010). Administration of testosterone increases the responsiveness of the amygdala to untrustworthy faces, perhaps due to heightened social vigilance (Bos et al., 2012a).

Testosterone decreases the amount of collaboration between two participants by increasing the egocentricity of the individual's choices (Wright, Bahrami, Johnson, Di Malta, Rees, Frith, et al., 2012), and decreases generosity (Zak, Kurzban, Ahmadi, Swerdloff, Park, Efremidze, et al., 2009). However, another study found that testosterone administration increases social cooperation in individuals with low levels of prenatal testosterone exposure (measured using 2D:4D ratio) (van Honk, Montoya, Bos, van Vugt, & Terburg, 2012), which provide some evidence that responses following testosterone administration may, in part, be dependent on early organizational effects.

Following testosterone (vs. placebo) administration, women have been found to increased activation in the thalamo-cingulate region, insula, and the cerebellum in response to infant crying, indicating testosterone may have a role in modulating parental care (Bos, Hermans, Montoya, Ramsey, & van Honk, 2010).

Testosterone also affects responsivity to reward. Using the IOWA gambling task, women show an increase in risk-taking after testosterone administration (van Honk, Schutter, Hermans, Putman, Tuiten, & Koppeschaar, 2004). Using a monetary incentive delay task, testosterone administration increases ventral striatum activation, associated with reward anticipation, in individuals with low appetitive motivation (behavior directed toward goals that are usually associated with reward processes) (Hermans, Bos, Ossewaarde, Ramsey, Fernandez, & van Honk, 2010).

Participants who believe that they received testosterone, regardless of whether they actually received it or not, behave more unfairly than those who believed that they were treated with placebo. In fact, testosterone administration increases the frequency of fair bargaining (Eisenegger, Naef, Snazzi, Heinrichs, & Fehr, 2010).

Although these studies provide interesting and novel evidence for testosterone administration effects, the sample sizes are small and further replication of the results is needed. These studies also include mainly females, and while they do control for the phase of the menstrual cycle, which itself predicts emotion recognition (Derntl, Kryspin-Exner, Fernbach, Moser, & Habel, 2008a) and brain function (e.g. amygdala response) (Derntl, Windischberger, Robinson, Lamplmayr, Kryspin-Exner, Gur, et al., 2008b), many of the women are also using oral contraceptives, which suppress ovarian hormone production (Fleischman, Navarrete, & Fessler, 2010). The effects of how all these factors interact and the effects of social and emotional behavior need further investigation.

Testosterone vs. oxytocin administration

Interestingly, another hormone oxytocin, has seemingly “opposite” results to those found for testosterone when examining its effect on aspects of human social behavior (Heinrichs, von Dawans, & Domes, 2009). In one study, duration and pattern of social gaze toward the eye region (predictive

of the ability to interpret the meaning of social situations and the intentions of others (Klin, Jones, Schultz, Volkmar, & Cohen, 2002)) in men was increased by administration of an intranasal dose of oxytocin (Guastella, Mitchell, & Dadds, 2008). Oxytocin increases trust in social situations, suggesting that it might serve an affiliative purpose in humans as well as animals (especially among in-group members) (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005).

Oxytocin exerts influence on neural circuits involved a wide range of social-cognitive abilities such as eye gaze, mentalizing, emotion-recognition and learning (Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008; Domes, Heinrichs, Glascher, Buchel, Braus, & Herpertz, 2007; Domes, Lischke, Berger, Grossmann, Hauenstein, Heinrichs, et al., 2010; Gamer, Zurowski, & Buchel, 2010; Kirsch, Esslinger, Chen, Mier, Lis, Siddhanti et al., 2005; Labuschagne, Phan, Wood, Angstadt, Chua, Heinrichs, et al., 2010; Pincus, Kose, Arana, Johnson, Morgan, Borckardt, et al., 2010; Riem, Bakermans-Kranenburg, Pieper, Tops, Boksem, Vermeiren, et al., 2011). Regions in these studies which are affected by oxytocin, such as the amygdala, fusiform gyrus, ventromedial prefrontal cortex, insula, superior temporal gyrus, and inferior frontal gyrus (Petrovic, Kalisch, Singer, & Dolan, 2008), are consistently atypical in conditions where difficulties in social cognition are a defining feature, such as ASC (Di Martino, Ross, Uddin, Sklar, Castellanos, & Milham, 2009; Lombardo, Baron-Cohen, Belmonte, & Chakrabarti, 2011). Extensive reviews on oxytocin effects can be found elsewhere (Heinrichs et al., 2009; Striepens, Kendrick, Maier, & Hurlmann, 2011).

The disparate effects of administering testosterone and oxytocin are becoming clearer, and it has recently been proposed that steroids and neuropeptides are important in different environments (Bos, et al., 2012b). For example, testosterone may increase vigilance and motivation for action and may reduce social cognition in environments that demand action (such as in emergencies or high stress situations). Neuropeptides such as oxytocin may increase social cognition in environments that are safe or that do not demand action. The subtleties of these interactions need further testing (Bos et al., 2012b). It will be important to consider the environment and situational contexts when interpreting the findings from research of this kind, where the vast majority of studies are conducted in laboratory settings.

Furthermore, unlike testosterone studies that mainly include females, the majority of studies of oxytocin have only included males. These samples are mainly chosen as a result of the practicalities of the side effects associated with each hormone. The generalizability of results from these studies has not yet been thoroughly tested, and possible sex-dependent outcomes have not been ruled out.

Future directions

In much of the research described in this chapter, the role of the social environment has not been considered in depth. Social interactions undoubtedly play an important role in the development of social and emotional behavior. For example, research on gender-based expectations may cause parents, teachers or caregivers to elicit and reinforce expected behavior from children (Stern & Karraker, 1989), thus shaping the child's behavior. Further work on the role of the environment and how various factors interact with hormone levels and behavior will be very important.

The relationships between hormones and behavior in humans are likely to depend on many factors and these studies in the main report correlations with hormone levels measured at a single time point. Research in animals has generally shown that hormonal effects on behavior may be dose and time-dependent (Cohen-Bendahan et al., 2005a; Hines, 2004), and these issues need to be clarified. The replication of results in larger sample sizes would also help to increase the range of hormone levels observed in these studies and assist in identifying any factors that are linked with levels in the extreme range.

It will also be valuable to further establish the relationships between direct measures of hormones (e.g. amniotic fluid or serum measures) and physical characteristics (e.g. 2D:4D ratio, or dermatoglyphics), which have been used as proxy measures of hormone exposure. The benefit of using these types of measurements is that they are easy to obtain and have also been linked to many areas of development. However, limited evidence exists for a relationship between these proxy measures and exposure to prenatal hormones (Lutchmaya, Baron-Cohen, Raggatt, Knickmeyer, & Manning, 2004). If such a link is further confirmed using direct measures of hormones, it could simplify future investigations of hormone effects.

In studies of puberty, it will be beneficial for the field to include in-depth studies that investigate the contribution of pubertal development, hormone levels, and social influences on development. The degree to which genetic variation is coupled with changes in hormone exposure is also unknown and it may be that changes in hormone levels are simply a manifestation of a genetic influence. This is an interesting area for future research, since investigations of current testosterone levels have shown rates of heritability between 50 and 66% (Harris, Vernon, & Boomsma, 1998; Hoekstra, Bartels, & Boomsma, 2006). Sex hormones also have an epigenetic role in changing gene expression throughout development and likely interact with sex chromosome effects on sexual differentiation (McCarthy & Arnold, 2011; McCarthy, Auger, Bale, De Vries, Dunn, Forger, et al., 2009), and further exploration of applications to social behavior would be important.

With regards to administration studies, it is worth reiterating the point made earlier, that the majority of studies that have used this methodology have restricted their samples to either a female sample when using testosterone, or a male sample when using oxytocin. As a result, the findings of the abovementioned administration studies may not necessarily generalize to samples of the opposite sex. Future studies should compare the responses of males and females to ascertain whether there may be any sex-dependent effects. Testosterone administration studies also include those who are using oral contraceptives, which itself is a hormone manipulation. It would be important for this area for studies to investigate how oral or hormone contraception may interact with the testosterone administration.

Conclusions

Research suggests that human social and emotional behavior, including theory of mind and empathy, are affected by gonadal hormones, in particular exposure to testosterone. The role of prenatal testosterone appears to be vital for early organization of the brain, and in the programming of sexual differentiation during critical periods of development. In humans, the most important period appears to be early-mid pregnancy. This finding has been repeated in studies looking at a number of behavioral measures and is also confirmed in those who are naturally exposed to elevated levels of the hormone through clinical conditions. It is also generally supported by studies in non-human mammals.

In later life, the effects of hormones such as testosterone during puberty have been shown to also predict behavior. It is thought that these effects “activate” or “fine-tune” the early organization of the brain, although the exact relationships between these two time periods are far from clear. To some extent, the activational effect of hormones during puberty appears to be dependent on exposure during the organizational period of early development, when key tissues are first formed.

Whilst the above conclusions appear to generally hold, there is still much work needed to further understand the subtle effects that specific changes to hormone levels may have. Administration of hormones to an individual can provide some further clues. Generally speaking, such studies have concluded that increased testosterone levels seem to be involved with decreased social and

emotional behavior, including theory of mind and empathy, whereas administration of oxytocin increases these social and emotional behaviors.

More recent studies are beginning to identify the physical processes that may be involved in the effects of hormones on development and behavior. This research is generally at an early stage, though there is an indication that specific areas of the brain are more developed in those with higher prenatal testosterone levels. Functional MRI (fMRI) studies involving administration of particular hormones also indicate greater or reduced response from specific brain regions due to changes in testosterone or oxytocin levels. Such experiments are useful because they do not require a longitudinal design, but at the same time cannot easily examine organizational effects. The ways in which steroids interact with neuropeptides and other hormones, as well as the cause of natural variation of sex steroids in general, is still not well understood.

The investigation of both organizational and activational hormone exposure on behavioral development remains an area needing much more detailed research. In addition to helping us map the process of human development, findings in this area could have major implications for clinical conditions characterized by social and emotional difficulties, such as autism.

The science examining hormonal effects on social and emotional development continues to evolve at a rapid pace. Many important studies are underway, including (in our lab in collaboration with the Danish State Serum Institute) a study testing if prenatal sex steroid hormones are elevated in a large sample of people who developed autism spectrum conditions, characterized in part by difficulties in understanding other minds.

References

- Abramovich, D. R. (1974). Human sexual differentiation—*in utero* influences. *Journal of Obstetrics and Gynecology* 81: 448–53.
- Abramovich, D. R., & Rowe, P. (1973). Foetal plasma testosterone levels at mid-pregnancy and at term: Relationship to foetal sex. *Journal of Endocrinology* 56: 621–2.
- Allison, C., Baron-Cohen, S., Wheelwright, S., Charman, T., Richler, J., Pasco, G., et al. (2008). The Q-CHAT (Quantitative Checklist for Autism in Toddlers): A normally distributed quantitative measure of autistic traits at 18–24 months of age: Preliminary report. *Journal of Autism and Developmental Disorders* 38: 1414–25.
- American Psychiatric Association. (1994). *DSM-IV Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington DC: APA.
- Arnold, A. P., & Gorski, R. A. (1984). Gonadal steroid induction of structural sex differences in the central nervous system. *Annual Review of Neuroscience* 7: 413–42.
- Auyeung, B., Ahluwalia, J., Thomson, L., Taylor, K., Hackett, G., O'Donnell, K. J., et al. (2013). Prenatal versus postnatal sex steroid hormone effects on autistic traits in children at 18–24 months of age. *Mol Autism*.
- Auyeung, B. (2008). *Foetal Testosterone, Cognitive Sex Differences and Autistic Traits*. Thesis, University of Cambridge, United Kingdom.
- Auyeung, B., Baron-Cohen, S., Ashwin, E., Knickmeyer, R., Taylor, K., Hackett, G., et al. (2009a). Fetal testosterone predicts sexually differentiated childhood behavior in girls and in boys. *Psychology Science* 20: 144–8.
- Auyeung, B., Baron-Cohen, S., Chapman, E., Knickmeyer, R., Taylor, K., & Hackett, G. (2009b). Fetal testosterone and autistic traits. *British Journal of Psychology* 100: 1–22.
- Auyeung, B., Baron-Cohen, S., Wheelwright, S., & Allison, C. (2008). The Autism Spectrum Quotient: Children's Version (AQ-Child). *Journal of Autism and Developmental Disorders* 38: 1230–40.
- Auyeung, B., Taylor, K., Hackett, G., & Baron-Cohen, S. (2010). Foetal testosterone and autistic traits in 18 to 24-month-old children. *Molecular Autism* 1: 11.

- Baron-Cohen, S. (1999). The extreme male-brain theory of autism. In H. Tager Flusberg (Ed.), *Neurodevelopmental Disorders* (pp. 401–29). Cambridge: MIT Press.
- Baron-Cohen, S., Lutchmaya, S., & Knickmeyer, R. (2004). *Prenatal Testosterone in Mind*. Cambridge: MIT Press.
- Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, L., & Lawson, J. (2001). Are intuitive physics and intuitive psychology independent? A test with children with Asperger syndrome. *Journal of Developmental and Learning Disorders* 5: 47–78.
- Baum, M. J., & Erskine, M. S. (1984). Effect of neonatal gonadectomy and administration of testosterone on coital masculinization in the ferret. *Endocrinology* 115: 2440–4.
- Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron*, 58, 639–650.
- Beck-Peccoz, P., Padmanabhan, V., Baggiani, A. M., Cortelazzi, D., Buscaglia, M., Medri, G., et al. (1991). Maturation of hypothalamic-pituitary-gonadal function in normal human fetuses: circulating levels of gonadotropins, their common alpha-subunit and free testosterone, and discrepancy between immunological and biological activities of circulating follicle-stimulating hormone. *Journal of Clinical Endocrinology and Metabolism* 73: 525–32.
- Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology* 42: 155–62.
- Bishop, D. V. M. (1998). Development of the children's communication checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. *Journal of Child Psychology and Psychiatry* 6: 879–91.
- Bos, P. A., Hermans, E. J., Montoya, E. R., Ramsey, N. F., & van Honk, J. (2010). Testosterone administration modulates neural responses to crying infants in young females. *Psychoneuroendocrinology* 35: 114–21.
- Bos, P. A., Hermans, E. J., Ramsey, N. F., & van Honk, J. (2012a). The neural mechanisms by which testosterone acts on interpersonal trust. *NeuroImage* 61: 730–7.
- Bos, P. A., Panksepp, J., Bluthé, R. M., & van Honk, J. (2012b). Acute effects of steroid hormones and neuropeptides on human social-emotional behavior: A review of single administration studies. *Frontiers in Neuroendocrinology* 33: 17–35.
- Bos, P. A., Terburg, D., & van Honk, J. (2010). Testosterone decreases trust in socially naive humans. *Proceedings of the National Academy of Sciences, USA* 107: 9991–5.
- Bramen, J. E., Hranilovich, J. A., Dahl, R. E., Chen, J., Rosso, C., Forbes, E. E., et al. (2012). Sex matters during adolescence: testosterone-related cortical thickness maturation differs between boys and girls. *PLoS One* 7: e33850.
- Breedlove, S. M. (1994). Sexual differentiation of the human nervous system. *Annual Review of Psychology* 45: 389–418.
- Chakrabarti, S., & Fombonne, E. (2005). Pervasive developmental disorders in preschool children: Confirmation of high prevalence. *American Journal of Psychiatry* 162: 1133–41.
- Chapman, E., Baron-Cohen, S., Auyeung, B., Knickmeyer, R., Taylor, K., & Hackett, G. (2006). Fetal testosterone and empathy: Evidence from the Empathy Quotient (EQ) and the “Reading the Mind in the Eyes” test. *Social Neuroscience* 1: 135–48.
- Christensen, L. W., & Gorski, R. A. (1978). Independent masculinization of neuroendocrine systems by intracerebral implants of testosterone or estradiol in the neonatal female rat. *Brain Research* 146: 325–40.
- Clements, J. A., Reyes, F. I., Winter, J. S., & Faiman, C. (1976). Studies on human sexual development. III. Fetal pituitary and serum, and amniotic fluid concentrations of LH, CG, and FSH. *Journal of Clinical Endocrinology and Metabolism* 42: 9–19.
- Cohen-Bendahan, C. C., van de Beek, C., & Berenbaum, S. A. (2005a). Prenatal sex hormone effects on child and adult sex-typed behavior: Methods and findings. *Neuroscience and Biobehavioral Reviews* 29: 353–84.

- Cohen-Bendahan, C. C. C., Buitelaar, J. K., van Goozen, S. H. M., Orlebeke, J. F., & Cohen-Kettenis, P. T. (2005b). Is there an effect of prenatal testosterone on aggression and other behavioral traits? A study comparing same-sex and opposite-sex twin girls. *Hormones and Behavior* 47(2): 230–7.
- Collaer, M. L., & Hines, M. (1995). Human behavioural sex differences: A role for gonadal hormones during early development? *Psychology Bulletin* 118: 55–107.
- Connellan, J., Baron-Cohen, S., Wheelwright, S., Batki, A., & Ahluwalia, J. (2000). Sex differences in human neonatal social perception. *Infant Behavior and Development* 23: 113–18.
- d’Ercole, C., Shojai, R., Desbriere, R., Chau, C., Bretelle, F., Piechon, L., et al. (2003). Prenatal screening: Invasive diagnostic approaches. *Child’s Nervous System* 19(7–8): 444–7.
- Derntl, B., Kryspin-Exner, I., Fernbach, E., Moser, E., & Habel, U. (2008a). Emotion recognition accuracy in healthy young females is associated with cycle phase. *Hormonal Behaviour* 53: 90–5.
- Derntl, B., Windischberger, C., Robinson, S., Lamplmayr, E., Kryspin-Exner, I., Gur, R. C., et al. (2008b). Facial emotion recognition and amygdala activation are associated with menstrual cycle phase. *Psychoneuroendocrinology* 33: 1031–40.
- Di Martino, A., Ross, K., Uddin, L. Q., Sklar, A. B., Castellanos, F. X., & Milham, M. P. (2009). Functional brain correlates of social and non-social processes in autism spectrum disorders: an activation likelihood estimation meta-analysis. *Biological Psychiatry* 65: 63–74.
- Domes, G., Heinrichs, M., Glascher, J., Buchel, C., Braus, D. F., & Herpertz, S. C. (2007). Oxytocin Attenuates Amygdala Responses to Emotional Faces Regardless of Valence. *Biological Psychiatry* 62: 1187–90.
- Domes, G., Lischke, A., Berger, C., Grossmann, A., Hauenstein, K., Heinrichs, M., et al. (2010). Effects of intranasal oxytocin on emotional face processing in women. *Psychoneuroendocrinology* 35: 83–93.
- Dumontheil, I., Burgess, P. W., & Blakemore, S. J. (2008). Development of rostral prefrontal cortex and cognitive and behavioural disorders. *Developmental Medicine & Child Neurology* 50: 168–81.
- Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., & Fehr, E. (2010). Prejudice and truth about the effect of testosterone on human bargaining behaviour. *Nature* 463: 356–9.
- Fausto-Sterling, A. (1992). *Myths of Gender*. New York: Basic Books.
- Fleischman, D. S., Navarrete, C. D., & Fessler, D. M. (2010). Oral contraceptives suppress ovarian hormone production. *Psychological Science* 21: 750–2; author reply 753.
- Forbes, E. E., & Dahl, R. E. (2010). Pubertal development and behavior: hormonal activation of social and motivational tendencies. *Brain Cognition* 72: 66–72.
- Gamer, M., Zurowski, B., & Buchel, C. (2010). Different amygdala subregions mediate valence-related and attentional effects of oxytocin in humans. *Proceedings of the National Academy of Sciences, USA* 107: 9400–5.
- Gillberg, C., Cederlund, M., Lamberg, K., & Zeijlon, L. (2006). Brief report: “the autism epidemic.” The registered prevalence of autism in a Swedish urban area. *Journal of Autism Development Disorders* 36: 429–35.
- Gotz, F., & Dorner, G. (1976). Sex hormone-dependent brain maturation and sexual behaviour in rats. *Endokrinologie* 68: 275–82.
- Goy, R. W., Bercovitch, F. B., & McBair, M. C. (1988). Behavioral masculinization is independent of genital masculinization in prenatally androgenized female rhesus macaques. *Hormones and Behavior* 22: 552–71.
- Goy, R. W., & McEwen, B. S. (1980). *Sexual Differentiation of the Brain*. Cambridge: MIT Press.
- Guastella, A. J., Mitchell, P. B., & Dadds, M. R. (2008). Oxytocin increases gaze to the eye region of human faces. *Biological Psychiatry* 63: 3–5.
- Hamilton, A., Plunkett, K., & Shafer, G. (2000). Infant vocabulary development assessed with a British Communicative Inventory: Lower scores in the UK than the USA. *Journal of Child Language* 27: 689–705.

- Hampson, E., Rovet, J. F., & Altmann, D. (1998). Spatial reasoning in children with congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Developmental Neuropsychology* 14: 299–320.
- Harris, J. A., Vernon, P. A., & Boomsma, D. I. (1998). The heritability of testosterone: A study of Dutch adolescent twins and their parents. *Behavior Genetics* 28: 165–71.
- Heinrichs, M., von Dawans, B., & Domes, G. (2009). Oxytocin, vasopressin, and human social behavior. *Frontiers in Neuroendocrinology* 30: 548–57.
- Hermans, E. J., Bos, P. A., Ossewaarde, L., Ramsey, N. F., Fernandez, G., & van Honk, J. (2010). Effects of exogenous testosterone on the ventral striatal BOLD response during reward anticipation in healthy women. *NeuroImage* 52: 277–83.
- Hermans, E. J., Putman, P., & van Honk, J. (2006). Testosterone administration reduces empathetic behavior: a facial mimicry study. *Psychoneuroendocrinology* 31: 859–66.
- Hermans, E. J., Ramsey, N. F., & van Honk, J. (2008). Exogenous testosterone enhances responsiveness to social threat in the neural circuitry of social aggression in humans. *Biological Psychiatry* 63: 263–70.
- Hines, M. (2004). *Brain Gender*. New York: Oxford University Press, Inc.
- Hines, M., Brook, C., & Conway, G. S. (2004). Androgen and psychosexual development: Core gender identity, sexual orientation and recalled childhood gender role behavior in women and men with congenital adrenal hyperplasia (CAH). *Journal of Sex Research* 41: 75–81.
- Hines, M., Davis, F. C., Coquelin, A., Goy, R. W., & Gorski, R. A. (1985). Sexually dimorphic regions in the medial preoptic area and the bed nucleus of the stria terminalis of the guinea pig brain: a description and an investigation of their relationship to gonadal steroids in adulthood. *Journal of Neuroscience* 5: 40–7.
- Hines, M., Fane, B. A., Pasterski, V. L., Matthews, G. A., Conway, G. S., & Brook, C. (2003). Spatial abilities following prenatal androgen abnormality: Targeting and mental rotations performance in individuals with congenital adrenal hyperplasia. *Psychoneuroendocrinology* 28: 1010–26.
- Hoekstra, R., Bartels, M., & Boomsma, D. I. (2006). Heritability of testosterone levels in 12-year-old twins and its relation to pubertal development. *Twin Research and Human Genetics* 9: 558–65.
- Hoffman, M. L. (1977). Sex differences in empathy and related behaviors. *Psychological Bulletin* 84: 712–22.
- Jost, A. (1970). Hormonal factors in the sex differentiation of the mammalian foetus. *Philosophical Transactions of the Royal Society of London: B Biological Sciences* 259: 119–30.
- Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., et al. (2005). Oxytocin modulates neural circuitry for social cognition and fear in humans. *Journal of Neuroscience* 25: 11489–93.
- Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry* 59: 809–16.
- Knickmeyer, R., Baron-Cohen, S., Raggatt, P., Taylor, K., & Hackett, G. (2006). Fetal testosterone and empathy. *Hormonal Behaviour* 49: 282–92.
- Knickmeyer, R. C., & Baron-Cohen, S. (2006). Fetal testosterone and sex differences. *Early Human Development* 82: 755–60.
- Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature* 435: 673–6.
- Labuschagne, I., Phan, K. L., Wood, A., Angstadt, M., Chua, P., Heinrichs, M., et al. (2010). Oxytocin attenuates amygdala reactivity to fear in generalized social anxiety disorder. *Neuropsychopharmacology* 35: 2403–13.
- Larsen, P. R., Kronenberg, H. M., Melmed, S., & Polonsky, K. S. (Eds). (2002). *Williams Textbook of Endocrinology*, 10th edn. Philadelphia: Saunders.
- Lombardo, M. V., Baron-Cohen, S., Belmonte, M. K., & Chakrabarti, B. (2011). Neural endophenotypes of social behaviour in autism spectrum conditions. In J. Decety & J. Cacioppo (Eds), *Oxford Handbook of Social Neuroscience* (pp. 830–47). Oxford: Oxford University Press.

- Lutchmaya, S., Baron-Cohen, S., & Raggatt, P. (2002a). Foetal testosterone and eye contact in 12 month old infants. *Infant Behavioural Development* 25: 327–35.
- Lutchmaya, S., Baron-Cohen, S., & Raggatt, P. (2002b). Foetal testosterone and vocabulary size in 18- and 24-month-old infants. *Infant Behavioural Development* 24: 418–24.
- Lutchmaya, S., Baron-Cohen, S., Raggatt, P., Knickmeyer, R., & Manning, J. T. (2004). 2nd to 4th digit ratios, fetal testosterone and estradiol. *Early Human Development* 77: 23–8.
- MacLusky, N., & Naftolin, F. (1981). Sexual differentiation of the central nervous system. *Science* 211: 1294–303.
- Martin, C. A., Kelly, T. H., Rayens, M. K., Brogli, B. R., Brenzel, A., Smith, W. J., et al. (2002). Sensation seeking, puberty, and nicotine, alcohol, and marijuana use in adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry* 41: 1495–502.
- McCarthy, M. M., & Arnold, A. P. (2011). Reframing sexual differentiation of the brain. *Nature Neuroscience* 14: 677–83.
- McCarthy, M. M., Auger, A. P., Bale, T. L., De Vries, G. J., Dunn, G. A., Forger, N. G., et al. (2009). The epigenetics of sex differences in the brain. *Journal of Neuroscience* 29: 12815–23.
- New, M. I. (1998). Diagnosis and management of congenital adrenal hyperplasia. *Annual Review of Medicine* 49: 311–28.
- Norman, R. J., Dewailly, D., Legro, R. S., & Hickey, T. E. (2007). Polycystic ovary syndrome. *Lancet* 370: 685–97.
- Palomba, S., Marotta, A., Di Cello, A., Russo, T., Falbo, A., Orio, F., et al. (2012). Pervasive developmental disorders in children of hyperandrogenic women with polycystic ovary syndrome: a longitudinal case-control study. *Clinical Endocrinology (Oxford)* 77(6): 898–904.
- Petersen, A. C., Crockett, L., Richards, M., & Boxer, A. (1988). A self-report measure of pubertal status: Reliability, validity, and initial norms. *Journal of Youth and Adolescence* 17: 117–33.
- Petrovic, P., Kalisch, R., Singer, T., & Dolan, R. J. (2008). Oxytocin attenuates affective evaluations of conditioned faces and amygdala activity. *Journal of Neuroscience* 28: 6607–15.
- Phoenix, C. H., Goy, R. W., Gerall, A. A., & Young, W. C. (1959). Organizing action of prenatally administered testosterone propionate on the tissues mediating mating behavior in the female guinea pig. *Endocrinology* 65: 369–82.
- Pincus, D., Kose, S., Arana, A., Johnson, K., Morgan, P. S., Borckardt, J., et al. (2010). Inverse effects of oxytocin on attributing mental activity to others in depressed and healthy subjects: a double-blind placebo controlled fMRI study. *Frontiers in Psychiatry* 1: 134.
- Quadagno, D. M., Briscoe, R., & Quadagno, J. S. (1977). Effects of perinatal gonadal hormones on selected non-sexual behavior patterns: A critical assessment of the non-human and human literature. *Psychological Bulletin* 84: 62–80.
- Resnick, S. M., Berenbaum, S. A., Gottesman, I. I., & Bouchard, T. J. (1986). Early hormonal influences on cognitive functioning in congenital adrenal hyperplasia. *Developmental Psychology* 22: 191–8.
- Reyes, F. I., Boroditsky, R. S., Winter, J. S., & Faiman, C. (1974). Studies on human sexual development. II. Fetal and maternal serum gonadotropin and sex steroid concentrations. *Journal of Clinical Endocrinology and Metabolism* 38: 612–17.
- Reyes, F. I., Winter, J. S., & Faiman, C. (1973). Studies on human sexual development. I. Fetal gonadal and adrenal sex steroids. *Journal of Clinical Endocrinology and Metabolism* 37: 74–8.
- Riem, M. M., Bakermans-Kranenburg, M. J., Pieper, S., Tops, M., Boksem, M. A., Vermeiren, R. R., et al. (2011). Oxytocin modulates amygdala, insula, and inferior frontal gyrus responses to infant crying: A randomized controlled trial. *Biological Psychiatry* 70(3): 291–7.
- Riem, M. M., Bakermans-Kranenburg, M. J., Pieper, S., Tops, M., Boksem, M. A., Vermeiren, R. R., et al. (2011). Oxytocin modulates amygdala, insula, and inferior frontal gyrus responses to infant crying: A randomized controlled trial. *Biological Psychiatry* 18: 663–7.
- Robinson, J., Judd, H., Young, P., Jones, D., & Yen, S. (1977). Amniotic fluid androgens and estrogens in midgestation. *Journal of Clinical Endocrinology* 45: 755–61.

- Romeo, R. D., Richardson, H. N., & Sisk, C. L. (2002). Puberty and the maturation of the male brain and sexual behavior: recasting a behavioral potential. *Neuroscience and Biobehaviour Review* 26: 381–91.
- Sangalli, M., Langdana, F., & Thurlow, C. (2004). Pregnancy loss rate following routine genetic amniocentesis at Wellington Hospital. *New Zealand Medical Journal* 117(1191): U818.
- Saxe, R. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. . In A. Leslie & T. German (Eds), *Handbook of Theory of Mind* Florence: Psychology Press .
- Schulz, K. M., Molenda-Figueira, H. A., & Sisk, C. L. (2009). Back to the future: The organizational-activational hypothesis adapted to puberty and adolescence. *Hormonal Behaviour* 55: 597–604.
- Scott, F. J., Baron-Cohen, S., Bolton, P., & Brayne, C. (2002). The CAST (Childhood Asperger Syndrome Test): Preliminary development of a UK screen for mainstream primary-school-age children. *Autism* 6: 9–13.
- Seron-Ferre, M., Ducsay, C. A., & Valenzuela, G. J. (1993). Circadian rhythms during pregnancy. *Endocrine Reviews* 14(5): 594–609.
- Smail, P. J., Reyes, F. I., Winter, J. S. D., & Faïman, C. (1981). The fetal hormonal environment and its effect on the morphogenesis of the genital system. In S. J. Kogan & E. S. E. Hafez (Eds), *Pediatric Andrology* (pp. 9–19). Boston: Martinus Nijhoff.
- Stern, M., & Karraker, K. H. (1989). Sex stereotyping of infants: A review of gender labeling studies. *Sex Roles* 20: 501–22.
- Striepens, N., Kendrick, K. M., Maier, W., & Hurlemann, R. (2011). Prosocial effects of oxytocin and clinical evidence for its therapeutic potential. *Frontiers in Neuroendocrinology* 32: 426–50.
- Tanner, J. M. (1962). *Growth at Adolescence: With a General Consideration of the Effects of Hereditary and Environmental Factors Upon Growth and Maturity From Birth to Maturity*, 2nd edn. Oxford: Blackwell.
- Terburg, D., Aarts, H., & van Honk, J. (2012). Testosterone affects gaze aversion from angry faces outside of conscious awareness. *Psychological Science* 23: 459–63.
- van Honk, J., Montoya, E. R., Bos, P. A., van Vugt, M., & Terburg, D. (2012). New evidence on testosterone and cooperation. *Nature* 485: E4–5; discussion E5–6.
- van Honk, J., & Schutter, D. J. (2007). Testosterone reduces conscious detection of signals serving social correction: implications for antisocial behavior.
- van Honk, J., Schutter, D. J., Bos, P. A., Kruijt, A. W., Lentjes, E. G., & Baron-Cohen, S. (2011). Testosterone administration impairs cognitive empathy in women depending on second-to-fourth digit ratio. *Proceedings of the National Academy of Science, USA* 108: 3448–52.
- van Honk, J., Schutter, D. J., Hermans, E. J., Putman, P., Tuiten, A., & Koppeschaar, H. (2004). Testosterone shifts the balance between sensitivity for punishment and reward in healthy young women. *Psychoneuroendocrinology* 29: 937–43.
- van Wingen, G., Mattern, C., Verkes, R. J., Buitelaar, J., & Fernandez, G. (2010). Testosterone reduces amygdala-orbitofrontal cortex coupling. *Psychoneuroendocrinology* 35: 105–13.
- Walsh, S. W., Ducsay, C. A., & Novy, M. J. (1984). Circadian hormonal interactions among the mother, fetus, and amniotic fluid. *American Journal of Obstetrics and Gynecology*, 150(6): 745–53.
- Williams, C. L., & Meck, W. H. (1991). The organizational effects of gonadal steroids on sexually dimorphic spatial ability. *Psychoneuroendocrinology* 16: 155–76.
- Williams, J., Scott, F., Stott, C., Allison, C., Bolton, P., Baron-Cohen, S., et al. (2005). The CAST (Childhood Asperger Syndrome Test): Test accuracy. *Autism* 9: 45–68.
- Wright, N. D., Bahrami, B., Johnson, E., Di Malta, G., Rees, G., Frith, C. D., et al. (2012). Testosterone disrupts human collaboration by increasing egocentric choices. *Proceedings of the Royal Society - Biological Sciences* 279: 2275–80.
- Zak, P. J., Kurzban, R., Ahmadi, S., Swerdloff, R. S., Park, J., Efremidze, L., et al. (2009). Testosterone administration decreases generosity in the ultimatum game. *PLoS One* 4: e8330.
- Zuckerman, M. (1971). Dimensions of sensation seeking. *Journal of Consulting and Clinical Psychology* 36: 45–52.

Understanding the genetics of empathy and the autistic spectrum

Bhismadev Chakrabarti and Simon Baron-Cohen

Understanding other minds is at the heart of social functioning. We constantly process a multitude of social cues across a range of sensory modalities, and respond to them. Empathy plays a central role in all such processes, and is defined as the capacity to understand the emotions and mental states of others, and respond to them with an appropriate emotion. There is considerable variation of empathy in the general population, and individuals with autism spectrum conditions (ASC) are largely represented at the low end of this distribution.

Recent years have seen significant advances in understanding the neurobiology of empathy and its individual differences (Chakrabarti & Baron-Cohen, 2006; Singer & Lamm, 2009). Independently, human molecular genetics has made enormous advances in the past decade, both in delineating the role of specific genes as well as making it possible to identify a large number of sequence variations (e.g. polymorphisms) in the whole human genome at once. This is not to discount the important role that experience and learning plays in the development of empathy, but this chapter focuses narrowly on the role of genes (Baron-Cohen, 2011; Bowlby, 1969). It is therefore timely to take a multilevel perspective in the study of empathy that spans from genes to cognition. In this chapter, we provide a brief overview of genetic approaches to study empathy and other trait measures of ASC. We then describe a recent study from our group, using dimensional phenotypic measures of empathy and autistic traits. Finally, we discuss some initial studies that relate genetic variation to “intermediate phenotypes” (also known as endophenotypes) relevant to autism and empathy.

Empathy and its heritability

Empathy is not a unitary construct, and most theoretical accounts suggest the existence of at least two factors, which are cognitive empathy (which includes “theory of mind”) and affective empathy (which includes “emotional contagion”). A third component that includes prosocial behaviour has also been suggested (Chakrabarti & Baron-Cohen 2006; Preston & de Waal, 2002). The importance of this fractionation is apparent in identifying neurological dissociations between the different components of empathy. For example, it is suggested that people with psychopathic personality disorder may have intact cognitive empathy (hence being able to deceive others), but impaired affective empathy (hence being able to hurt others), whilst people with autism may show the opposite profile (hence finding the social world confusing because of their deficit in cognitive empathy, but not being over-represented among criminal offenders, having no wish to hurt others, suggesting their affective empathy may be intact; Baron-Cohen, 2011; Jones, Happé, Gilbert, Burnett, & Viding, 2010; Rogers, Viding, Blair, Frith, & Happé, 2006).

Before embarking on a discussion about the genetic underpinnings of empathy, it is essential to establish that an individual's genetic composition contributes to his/her levels of empathy. A standard approach to do this has been to test for heritability of "trait empathy" (i.e. stable individual differences in empathy) or other aspects of social behaviour by comparing monozygotic (MZ) and dizygotic (DZ) twins. Nearly all of these studies have shown a greater correlation of empathy measures in MZ compared with DZ twins, suggesting a partially genetic basis for trait empathy (Davis, Luce, & Kraus, 1994; Loehlin & Nichols, 1976; Matthews, Batson, Horn, & Rosenman, 1981). Measures have included the Questionnaire Measure of Emotional Empathy (QMEE) (Mehrabian & Epstein, 1972). (Rushton, Fulker, Neale, Nias, & Eysenck, 1986), in a large-scale twin study, and which suggested a high heritability estimate of 68% for emotional empathy. Other twin studies, particularly in children, have used behavioural observation paradigms of empathy in a laboratory situation. These involve simulating scripted situations (e.g. the experimenter tripping on a chair, or the mother of the child getting her finger caught while closing a suitcase), while video-recording the child's reactions. A study of 14- and 20-month-old twins using this paradigm confirmed a genetic contribution to empathic concern (Zahn-Waxler, Radke-Yarrow, Wagner, & Chapman, 1992). A more recent twin study on 409 twin pairs by the same group showed that genetic effects on the prosocial behaviour component of empathy (measured using video-recorded behaviour in a laboratory setting) increase with age, while shared environmental effects decrease with age (Knafo, Zahn-Waxler, Van Hulle, Robinson, & Rhee, 2008). In contrast, twin studies of cognitive empathy (measured using a theory of mind paradigm) have reported a greater genetic component in early compared with late childhood (Hughes, Jaffi, Happé, Taylor, Caspi, & Moffitt, 2005). In summary, there is considerable evidence for a moderate to high genetic contribution to each of the component processes of empathy, quantified using observational measures.

In adults, self-reported measures of empathy have been widely used as one of the key measures of social behaviour. A number of such trait and observational measures of social behaviour have been studied for genetic contributions (Ebstein, Israel, Chew, Zhong, & Knafo, 2010). Important among these are behavioural assays of face perception and emotion perception. Face recognition is associated with a strong genetic component (Wilmer et al., 2010). Recognition of emotions from the eye region of the face, as tested by the "Reading the Mind in the Eyes" Task (RMET), shows a strong degree of familiarity (Baron-Cohen & Hammer, 1997; Losh & Piven, 2007;). Questionnaire measures of social functioning using the Social Responsivity Scale (SRS; Constantino & Todd, 2000, 2005; Sung, Dawson, Munson, Estes, Schellenberg, & Wijsman, 2005) and of autistic traits using the Autism Spectrum Quotient (AQ; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) also reveal strong familiarity (Bishop, Maybery, Maley, Wong, Hill, & Hallmayer, 2004; Wheelwright, Auyeung, Allison, & Baron-Cohen, 2010), as well as heritability in twin studies (Hoekstra, Bartels, Verweij, & Boomsma, 2007). These studies corroborate findings from the early twin studies in suggesting a genetic underpinning for empathy and social behaviour relevant to ASC.

Insights from autism genetics

ASC entail a disability in social and communication development, alongside unusually narrow interests ("obsessions") and repetitive behaviour (APA, 1987; ICD-10, 1994). ASC have a genetic basis, indicated by significantly higher concordance rates in MZ than in DZ twins, and with some heritability estimates of over 90% (Bailey, Le Couteur, GottesmanBolton, Simmonoff, Yuzda, et al., 1995; Folstein & Rutter, 1977). Over the last three decades, a number of strategies have been used to discover genes related to ASC. A common feature in most of these studies has been

the use of clinical diagnosis of ASC as a categorical phenotype. In these studies, people with a diagnosis of ASC are compared with a group of people without a clinical diagnosis, matched on a variety of measures. This approach has implicated multiple genes, along with environmental (Wagner, Reuhl, Cheh, McRae, & Halladay, 2006) and epigenetic factors (Crespi & Badcock, 2008; Nagarajan Patzel, Martin, Yasui, Swanberg, & Hertz-Picciotto, 2008; LaSalle & Yasui, 2009). Mixed evidence from genome-wide linkage studies of samples that do not differentiate between classic (low-functioning) autism and Asperger syndrome (AS) has found linkage peaks in nearly all chromosomes (Abrahams & Geschwind, 2008).

Genome wide association studies (GWAS) are a more recent development, and use oligonucleotide microarrays that allow for simultaneous genotyping of common polymorphisms from nearly all known human genes. The initial GWAS on autism, using the traditional case-control design, found significantly associated polymorphisms in genes located on multiple chromosomes (AGPC, 2007; Wang, Zhang, Ma, Bucan, Glessner, Abrahams, et al., 2009). Oligonucleotide microarrays enable the detection of single nucleotide variations (e.g. change from an A to a C). Advances in the last five years have allowed the detection of larger segments of DNA across the genome (usually 1000 bases or longer), which are present in multiple copies or are deleted altogether in certain individuals. These are referred to as copy-number variations (CNV), and are believed to arise as *de-novo* events during gametogenesis. Rare *de novo* copy number variations (CNV) can potentially account for up to 10–24% of cases in families have only one child with ASC (Jacquemont, Sanlaville, Redon, Raoul, Cormier-Daire, Lyonnet, et al., 2006; Pinto, Pagnamenta, Klei, Anney, Merico, Regan, et al., 2010; Sebat, Lakshmi, Malhotra, Troke, Lese-Martin, Walsh, et al., 2007). In summary, case-control genetic studies of ASC suggest that:

1. ASC is an oligogenic condition (i.e. it is unlikely that there will be a single gene whose malfunction will explain all features of this condition).
2. Both rare CNVs as well as common sequence variants (single nucleotide polymorphisms; SNPs) are associated with this condition (Arking, Cutler, Brune, Teslovich, West, Ikeda, et al., 2008; Corvin, Craddock, & Sullivan, 2010; Glessner, Wang, Cai, Korvatska, Kim, Wood, et al., 2009; Holt & Monaco, 2011; Pinto et al., 2010; Wang et al., 2009).

While genotyping common and rare sequence variants of the whole human genome has become a routine procedure over the last few years, most studies have continued to use the classic case-control design. This poses some potential problems, particularly for autism research. The heterogeneity within ASC is not captured in this design, as most of these studies group people with classic autism together with those on the broader spectrum (having a diagnosis of high functioning autism (HFA) or AS). This raises the possibility of potential confounds due to factors such as language delay, below average IQ (seen in classic autism, but not in AS) or co-occurring (a term we prefer to the more medical term “comorbid,” for obvious reasons) conditions such as epilepsy and hyperactivity. In addition, a commonly used measure for verifying a current clinical diagnosis of autism (e.g. the Autism Diagnostic Observation Schedule (ADOS) (Lord, Rutter, Goode, Heemsbergen, Jordan, Mawhood, et al., 1989) is

1. optimized for diagnosing classic autism, and not AS/HFA;
2. does not include one key dimension of the autistic symptomatology (repetitive behaviour) in its final scoring algorithm, both of which could result in a biased sampling within the clinical cohorts.

In view of the heterogeneity within ASC, and given the existence of the “broader autism phenotype” (BAP) (Piven, Palmer, Jacobi, Childress, & Arndt, 1997) or subthreshold instances of ASC, an emerging consensus in autism phenotypic studies suggests that autistic traits are distributed on a continuum not just within clinic samples, but right across the general population. Behavioural genetic studies confirm this, suggesting that the etiology of autistic traits is similar in the general population as well as the extreme ends of the continuum (Robinson, Koenen, McCormick, Munir, Hallett, Happé, et al., 2011). The AQ is one such trait measure that captures the population variability in autistic traits in both social and repetitive behaviour domains (Baron-Cohen et al., 2001). Another self-report measure focusing specifically on empathy is the empathy quotient (EQ), a 40-item questionnaire that provides a continuous range of scores across the general population (Baron-Cohen & Wheelwright, 2004). These, and other similar trait measures such as the SRS (Constantino, Przybeck, Friesen, & Todd, 2000) provide a dimensional measure of the social functioning in the general population, and people with ASC tend to cluster toward the low end of the score distribution.

Bridging the genotype-phenotype gap in ASC using a dimensional and case-control approach

Interestingly, while most phenotypic studies of ASC (using questionnaires, computer-based tasks, and neuroimaging) have focused on the higher functioning end of the autistic spectrum, large-scale genetic studies have primarily tested the “lower-functioning” end, largely focusing on classic autism. This presents a disconnection between advances at the phenotypic and genotypic ends of the sequence from DNA to cognition. A small number of pioneering studies have attempted to bridge this disconnect by studying the dimensional phenotypes within ASC using linkage and association studies (Campbell, Warren, Sutcliffe, Lee, & Levitt, 2010; Conciatori, Stodgell, Hyman, O’Bara, Militerni, Bravaccio, et al., 2004; Losh, Sullivan, Trembath, & Piven, 2008). We attempted to bridge this disconnect by conducting two parallel candidate gene association studies, which we describe in the next section. The first is of empathy (measured using the EQ) and autistic traits (measured using the AQ) in the general population. The second is of Asperger syndrome, which is marked by social and behavioural impairments and unusually narrow interests, but is not associated with language or general cognitive delays during development.

A key feature of our studies was in the choice of multiple candidate genes from three groups of genes, defined by gene function. This approach has been used in other conditions (Pharoah, Tyrer, Dunning, Easton, & Ponder, 2007), but not in the study of ASC. Traditionally, genetic association studies of ASC have either studied one or a small number of candidate genes, or on the whole genome (Losh et al., 2008). We chose 68 candidate genes for these two experiments, derived from three functional categories:

1. Sex hormone-related genes;
2. Genes involved in neural development and connectivity;
3. Genes involved in social and emotional responsivity (see Table 18.1). We searched for common genetic variants (SNPs) on the assumption that autistic traits are continuously distributed in the general population so the genetic contributions to individual differences in empathy or autistic traits are likely to be normative variants rather than “disease”-causing mutations.

Each of the three functional categories derives from a clear neurocognitive theory of ASC, outlined next.

The fetal androgen theory suggests that genes involved in sex steroid synthesis and transport might be related to empathy and ASC (Baron-Cohen, Knickmeyer, & Belmonte, 2005; Auyeung et al., chapter 17 in this book). Much of the empirical basis of this theory derives from studies that have measured levels of fetal testosterone (FT), measured in amniotic fluid in the general population. FT levels correlate *negatively* with markers of social behaviour, such as eye-contact at 12 months old, vocabulary size at 24 months old (Lutchmaya, Baron-Cohen, & Raggatt, 2002), and social development at 4 years old (Knickmeyer, Baron-Cohen, Raggatt, & Taylor, 2005). FT correlates negatively with scores on the EQ and the RMET at 8 years old (Chapman, Baron-Cohen, Auyeung, Knickmeyer, Hackett, & Taylor, 2006). FT levels also correlate *positively* with narrow interests at 4 years old (Knickmeyer et al., 2005), SQ, AQ at 8 years old (Auyeung, Baron-Cohen, Chapman, Knickmeyer, Taylor, & Hackett, 2006; Auyeung, Baron-Cohen, Ashwin, Knickmeyer, Taylor, Hackett, et al., 2009), and autistic traits at as young as 18–30 months of age (Auyeung, Taylor, Hackett, & Baron-Cohen, 2010).

The neural connectivity theory, based on evidence from studies of rodent and human brains, suggest that the key abnormality in autism might be related to neural growth and connectivity (Belmonte, Cook Jr, Anderson, Rubenstein, Greenough, Beckel-Mitchener, et al., 2004; Wass, 2011). ASC has a neurodevelopmental origin and an emerging body of genetic evidence suggests a crucial role for genes involved in neural growth, synaptic development and function (Bourgeron, 2009). At the phenotypic end, several studies show functional (Just, Cherkassky, Keller, & Minshew, 2004; Minshew & Williams, 2007; Shih, Shen, Öttl, Keehn, Gaffrey, & Müller, 2010; Villalobos, Mizuno, Dahl, Kemmotsu, & Müller, 2005; Welchew, 2005) and structural underconnectivity in the autistic brain (Barnea-Goraly, Kwon, Menon, Eliez, Lotspeich, & Reiss, 2004; Keller, Kana, & Just, 2007; Sahyoun, Belliveau, & Mody, 2010; Sundaram, Kumar, Makki, Behen, Chugani, & Chugani, 2008), which is also marked by abnormal growth patterns (Courchesne, Pierce, Schumann, Redcay, Buckwalter, Kennedy, et al., 2007; Courchesne, Campbell & Solso, 2011). We therefore hypothesized that variations in genes governing neural development and synaptic function could contribute to autistic traits.

Finally, the social-emotional responsivity theory suggests that the atypical social behaviour patterns in ASC might be related in part to genes known to modulate social behaviour in animals (Chakrabarti, Kent, Suckling, Bullmore, & Baron-Cohen, 2006; Dawson, Carver, Meltzoff, Panagiotides, McPartland, & Webb, 2002; Insel, O'Brien, & Leckman, 1999). These included genes involved in the oxytocin and vasopressin systems, as well as other neuropeptides involved in endogenous reward systems, such as opioids and cannabinoids. Some of these genes have been associated with autism in previous genetic studies, and these are shown in Table 18.1.

These 68 candidate genes were tested in two experiments. 216 SNPs with a minor allele frequency (MAF) ≥ 0.2 in the Caucasian population were chosen from these genes (full list of SNPs are available in (Chakrabarti, Dudbridge, Kent, Wheelwright, Hill-Cawthorne, Allison, et al., 2009)). This approach, of selecting multiple common SNPs per gene, has the advantage of checking for informative associations both directly and indirectly (Collins, Guyer, & Chakravarti, 1997). The median SNP density across all genes was one SNP per 14.1 kb. 125 of these SNPs have been genotyped in one or more populations in the HapMap database (Release 23a). All volunteers contributed mouth swabs for DNA extraction. These were anonymized and DNA was genotyped for the 216 SNPs using standard PCR-based assays (TaqMan® SNP genotyping assays, Applied Biosystems Inc., California, USA). The two experiments conducted were as follows:

1. *Experiment 1:* An association study for EQ and AQ was conducted on the population sample ($n = 349$) using non-parametric analysis of variance for each SNP. Chi-square statistics and

Table 18.1 List of all genes included in the association study, along with brief functional roles where known. Genes marked in bold indicate those previously linked to ASC through genetic linkage/association studies

Neural development and connectivity	
<i>NGF, BDNF, NTF3, NTF5, NGFR, NTRK1, NTRK2, NTRK3, TAC1, IGF1, IGF2</i>	Neuronal survival, differentiation and growth.
<i>RAPGEF4</i>	Growth and differentiation of neurons. Mutations associated with classic autism.
<i>VGF</i>	Upregulated directly by NGF and expressed in neuroendocrine cells.
<i>VEGF</i>	Promotes cell growth and migration, especially during angiogenesis and vasculogenesis, often observed during hypoxia. Modulated directly by PTEN.
<i>ARNT2</i>	Neural response to hypoxia
<i>NLGN1, NLGN4X, AGRIN</i>	Synapse formation and maintenance in CNS neurons. <i>NLGN4X</i> mutations have been linked to autism.
<i>NRCAM</i>	Neuronal adhesion and directional signalling during axonal cone growth.
<i>EN-2(AUTS1)</i>	Neuronal migration and cerebellar development. <i>EN-2</i> has been previously linked to ASCs in several studies.
<i>HOXA1</i>	Hindbrain patterning. Mixed evidence suggests a link with ASCs.
Social and emotional responsivity	
<i>OXT, OXTR, AVPR1A, AVPR1B</i>	Linked to social attachment behaviour in humans and other mammals. <i>AVPR1A</i> and <i>OXTR</i> have previously been associated with ASCs.
<i>CNR1, OPRM1, TRPV1</i>	Mediate endogenous reward circuits, in tandem with dopaminergic pathways. Implicated in underlying rewarding features of social interactions.
<i>MAOB</i>	Synaptic breakdown of dopamine and serotonin. Suggested links with social cognition.
<i>WFS1</i>	Mutations linked to affective disorders. Overexpressed in amygdala during fear response, though exact functional role is not known.
<i>GABRB3, GABRG3, GABRA6, ABAT</i>	Mediate inhibitory (GABA-ergic) neurotransmission as well as play a role in early cortical development. <i>GABRA6</i> is expressed strongly in the cerebellum; <i>GABRB3, GABRG3, ABAT</i> have all been associated with ASCs.
<i>VIPR1</i>	Suggested involvement in neural pathways underlying pheromone processing. Mutations associated with social behavioural abnormalities in mice. Its endogenous ligand (VIP) shows an overexpression in neonatal children with autism.
Sex hormone biosynthesis, metabolism and transport	
<i>DHCR7</i>	Metabolism of cholesterol: precursor for sex hormones (mutations associated with near-universal presence of ASC)

(Continued)

Table 18.1 (Continued)

Neural development and connectivity	
<i>CYP11A1, CYP11B1, CYP3A, CYP7A1, CYP11A, CYP11B1, CYP17A1, CYP19A1, CYP21A2, POR</i>	Synthesis of sex hormones such as progesterone, estrogen, cortisol, aldosterone and testosterone. <i>CYP21A2</i> and <i>POR</i> mutations associated with CAH.
<i>HSD11B1, HSD17B2, HSD17B3, HSD17B4</i>	Local regulation of sex steroids.
<i>STS, SULT2A1, SRD5A1, SRD5A2</i>	Steroid hormone metabolism
<i>SHBG, SCP2, TSPO, SLC25A12, SLC25A13</i>	Intracellular transport of sex steroids as well as their important precursors and/or metabolites. Mixed evidence suggests an association of <i>SLC25A12</i> with classic autism.
<i>AR</i>	Intracellular receptor for testosterone
<i>ESR1, ESR2</i>	Receptors for estrogen
<i>CGA, CGRPR, LHB, LHRHR, LHCGR, FSHB</i>	Regulation of reproductive functions.

asymptotic *p*-values (two-tailed) were generated from this test. A sex-specific analysis was conducted for all X-linked genes.

2. *Experiment 2:* A case-control association study of AS was conducted on all cases of AS (*n* = 174) and a subset of the population sample (*n* = 155). The controls were selected to be sex-matched with the cases, whilst having an AQ score <25. An AQ <25 cut-off was employed to exclude a small number of individuals who scored high on AQ even though they did not have a formal diagnosis. For each SNP, a Cochran-Armitage chi-square statistic (1 d.f.) was calculated to test the null hypothesis that the different alleles have the same distribution in cases and controls. Asymptotic *P*-values (two-tailed) were calculated.

To control for multiple testing of SNPs within genes as well as for multiple phenotypes, permutation testing was conducted using UNPHASED (Dudbridge, 2008) for Experiment 1, and using PLINK (Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, et al., 2007) for Experiment 2. Since each candidate gene was individually selected on the basis of a priori hypothesis, independent of other genes, permutation tests were performed separately for each gene. In each permutation, the phenotypes were randomly reassigned among participants, keeping the genotypes fixed to preserve their correlation structure. The multiple phenotypes for each subject were permuted together so as to preserve the correlation structure among phenotypes. Each SNP was then tested for association to each permuted phenotype and the minimum *P*-value recorded. The permutation was repeated 1000 times and the corrected *P*-value was the estimated proportion of permutations in which the minimum *P*-value was less than or equal to the minimum *P*-value seen in the original data. When the Family Wise Error Rate (FWER)- corrected *P*-value is significant, we may infer that at least one SNP in the gene is associated and that there is gene-wise significance. This gene-wise *p*-value thus reflects the *p*-value of the most significant SNP after FWER correction.

In Experiment 1, autistic traits and/or empathy (measured on AQ and/or EQ) were nominally associated at *P* < = 0.05 with SNPs from 19 genes. In Experiment 2, SNPs from 14 genes were nominally associated at *P* < = 0.05 with AS. Across both experiments, six genes showed nominal

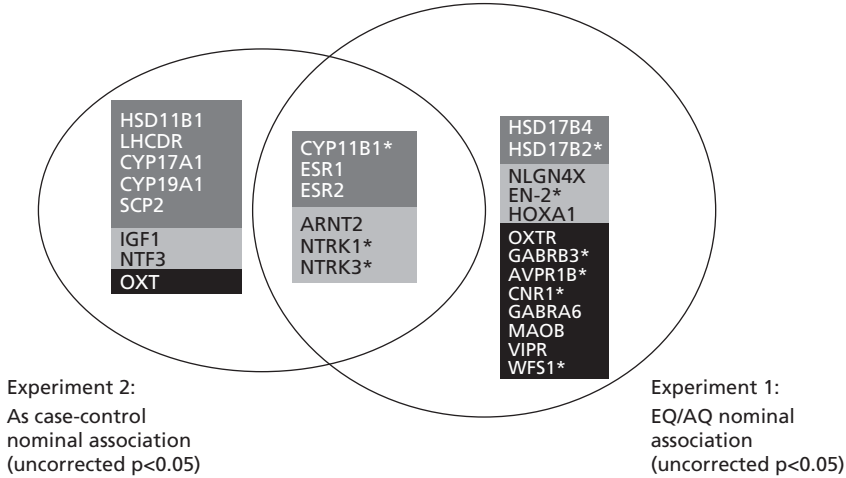


Figure 18.1 Genes showing nominal association with (1) AS case-control analysis, (2) autistic trait measures (AQ, EQ) in the population sample. Intersections summarize genes that show a nominal association in both experiments. Gene functional groups are colour coded: gray (sex hormone-related), light gray (neural connectivity related) and dark gray (social-emotional responsiveness related). Genes in bold indicate replications of associations reported in earlier studies. * Indicates a nominally significant association with EQ. See also Plate 11.

significance at $P \leq 0.05$. (See Figure 18.1 for a distribution of all nominally significant genes across the two experiments).

Eight genes in Experiment 1 and 5 genes in Experiment 2, showed gene-wise significance after 1000 permutations across all phenotypes. Two genes (*CYP11B1* and *NTRK1*) survived FWER correction in both the experiments, and are therefore strong candidates for future replication studies. Genes in all three functional groups were found to be significantly associated both with empathy and/or autistic traits, as well as with a diagnosis of AS. This provides further support for the non-unitary nature of autistic traits and AS (Happé, Ronald, & Plomin, 2006).

In the sex steroid group, the estrogen receptor beta (*ESR2*) was associated significantly in both experiments. Particularly, the C allele in rs1271572 and rs1152582 was associated with higher AQ in the typical population, and were also found to be more frequent in cases than in controls. *ESR2* codes for the main estrogen receptor expressed in the brain. In the fetal brain testosterone is aromatized to estradiol and exerts its effects on neural development through acting on these receptors, and mediating selective cell survival. It promotes the defeminization of the developing male brain in mice (Kudwa, Bodo, Gustafsson, & Rissman, 2005). Estrogen is thought to mediate social interaction in rodents, and this is supported by the presence of estrogen receptors in areas of the brain involved in emotion and affective behaviour, such as the amygdala and the hippocampus.

CYP17A1 catalyses the production of dehydroepiandrosterone (DHEA, a precursor of testosterone), as well as androstenedione (a precursor of estradiol). Higher levels of androstenedione were found in males and females with ASC in a recent study on an independent sample (Ruta, Ingudomnukul, Taylor, Chakrabarti, & Baron-Cohen, 2011). Polymorphisms in *CYP17A1* have been associated with PCOS in women (Park, Lee, Ramakrishna, Cha, & Baek, 2008), a condition known to be elevated in ASC (Ingudomnukul, Baron-Cohen, Wheelwright, & Knickmeyer, 2007). *CYP11B1* is cellularly localized in the mitochondria and converts 11-deoxycortisol to cortisol.

Polymorphisms in this gene and the *CYP11A* gene are associated with congenital adrenal hyperplasia (CAH; Kuribayashi, Nomoto, Massa, Oostdijk, Wit, Wolffenbuttel, et al., 2005) in which FT is elevated. CAH is associated with higher AQ than in the general population (Knickmeyer, Baron-Cohen, Fane, Wheelwright, Mathews, Conway, et al., 2006). Together, these results implicate genes involved in the synthesis and metabolism of sex steroids in the aetiology of autistic traits, empathy, and AS, and provides some of the first genetic evidence in support of the role of sex-steroids in ASC and related trait measures.

In the neurodevelopmental group, four genes (*HOXA1*, *NLGN4X*, *NTRK1*, and *ARNT2*) survived FWER correction. rs10951154 in *HOXA1* has been previously associated with head size in ASC, as well as with head growth rate (Muscarella, Guarnieri, Sacco, Militeri, Bravaccio, Trillo, et al., 2007). We found that the G-allele carriers had more autistic traits than the AA homozygotes. This is consistent with the finding that the G allele has been found to be associated with larger head size and greater head growth rate (Muscarella et al., 2007). rs12836764 in the *NLGN4X* UTR was significantly associated with both EQ and AQ in females. This supports earlier findings implicating this gene in autism (Jamain, Quach, Betancur, Råstam, Colineaux, Gillberg, et al., 2003). A large-scale association study of autism found a significant association with neurexins (AGPC, 2007) that interact with neuroligins in mediating glutamatergic synaptogenesis. Among the molecules related to neurotrophin function, a strong association was seen in *NTRK1* with empathy (in Experiment 1), and with AS (in Experiment 2). *NTRK1* is situated within a peak (1q21–2) reported in the first ever linkage study of AS (Ylisaukko-oja, Nieminen-von Wendt, Kempas, Sarenius, Varilo, von Wendt, et al., 2004) and thus provides an independent validation. Nerve growth factor (NGF), signaling through TrkA (the protein product of *NTRK1*), mediates most neurotrophic action of NGF (Sofroniew, Howe, & Mobley, 2001). A primary role of the TrkA in the developing brain is in determining the fate and growth of neurites, in whether they become axons or dendrites (Da Silva, Hasegawa, Miyagi, Dotti, & Abad-Rodriguez, 2005). Additionally, two SNPs in the *ARNT2* gene were found to be associated in both the experiments. This gene is involved both in the development of the neuroendocrine cells in the hypothalamus (Michaud, DeRossi, May, Holdener, & Fan, 2000), as well as in the neural response to hypoxia (Maltepe, Keith, Arsham, Brorson, & Simon, 2000). These findings point to a key role played by these neurodevelopmental genes in the development of empathy and autistic traits.

In the social-emotional responsivity group, four genes (*MAOB*, *GABRB3*, *WFS1*, *OXT*) were found to be significant after FWER correction. *MAOB* was significantly associated in females only, and this is consistent with the earlier studies showing the importance of this locus in social cognition, both in humans and mouse models (Good, Lawrence, Simon-Thomas, Price, Ashburner, Friston, et al., 2003; Grimsby, Toth, Chen, Kumazawa, Klaidman, Adams, et al., 1997). The rationale for testing GABA-related genes came from the fact that social behaviour has been linked to GABA-ergic activity in the CNS (File & Seth, 2003), and that GABA receptors play a crucial role early in cortical development through their effect on neuronal migration, as well as on development of excitatory and inhibitory synapses. In this sense, GABA-related genes could have been placed in both the neurodevelopmental group of candidate genes too. We found *GABRB3* was significantly associated with empathy (EQ) in the typical sample, thus corroborating a role of this locus (15q11-q13) in autism (Ashley-Koch, Mei, Jaworski, Ma, Ritchie, Menold, et al., 2006; Buxbaum, Silverman, Smith, Greenberg, Kilfarski, Reichart, et al., 2002). *Gabrb3* knockout mice have been shown to demonstrate low social and exploratory behaviour as well as smaller cerebellar vermal volumes, pointing to a potential animal model for autism (DeLorey, Sahbaie, Hashemi, Homanics, & Clark, 2007).

Another significant association in this functional group of genes was the Wolframin (*WFS1*) gene. Wolframin is strongly expressed in the amygdala, especially in response to fear-inducing

stimuli (Koks, Planken, Luuk, & Vasar, 2002). The amygdala is one of the key brain regions where functional and structural abnormalities have been consistently found in ASC (Baron-Cohen, Ring, Bullmore, Wheelwright, S., Ashwin, & Williams, 2000). 2 SNPs in *WFS1* showed a strong association with both AQ and EQ. One of these, rs734312, is a non-synonymous coding SNP and belongs to a haplotype that shows an increased risk for affective disorders (Koido, Kōks, Nikopensius, Maron, Altmäe, Heinaste, et al., 2004). Finally, three genes from the oxytocin-vasopressin system (*OXTR*, *OXT*, and *AVPR1B*) were found to be nominally associated with ASC and/or with AQ and EQ. These genes have suggestive links with autism (Insel et al., 1999; Jacob, Brune, Carter, Leventhal, Lord, & Cook Jr, et al., 2007; Tops et al., 2011; Wermter, Kamp-Becker, Hesse, Schulte-Körne, Strauch, K., & Remschmidt, 2009; Wu Jia, Ruan, Liu, Guo, Shuang, et al., 2005; Wu et al., 2012) and with social behaviour in animal models. Of these, *OXT* survived a FWER correction in Experiment 2. Oxytocin is of particular interest, given the recent reports of oxytocin levels being low in autism, and treatment effects of both intranasal and intravenous administration of oxytocin (Hollander, Novotny, Hanratty, Yaffe, DeCaria, Aronowitz, et al., 2003). Oxytocin levels are also correlated with empathy and prosocial measures, such as the Eyes Test (Domes, Heinrichs, Michel, Berger, & Herpertz, 2007) and trust in neuroeconomics (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). This provides partial support for the involvement of the oxytocin-vasopressin system in autistic traits. Together, these results support the idea that genes implicated in social and emotional responsivity contribute to individual differences in traits related to ASC.

In summary, in the two studies described above, we identified 9 candidate genes, some of which are associated with autistic traits in the general population and/or AS. These genes fall into the three functional categories related to sex-steroid synthesis and metabolism, neural development and connectivity, and social-emotional responsivity, providing some support for three theories of autism. It is essential that these are replicated in independent samples, and validated through molecular genetic techniques such as gene expression measurement. Importantly, these associations should be validated against other relevant endophenotypes.

Endophenotypes and future directions

Endophenotypes are defined as measurable intermediate phenotypes that are generally closer to the action of the gene and thus exhibit higher genetic signal-to-noise ratios (Gottesman & Gould, 2003). A range of endophenotypic measures have been suggested for empathy and autistic behaviour, and social cognition and emotion processing ranks highly among these (Losh & Piven, 2007). In our study described above, we did a preliminary test of two such endophenotypic measures (the “Reading the Mind in the Eyes” Test, and the Embedded Figures Test) for cross-validation of our trait association results, in a small subset of the general population sample. This found a nominal association in seven genes with these measures that overlapped with the significantly associated genes in either/both of the two main experiments (Chakrabarti et al., 2009). While this analysis was preliminary, and under-powered, this provides a framework for future studies. Additional endophenotypes that have been put forward to study social behaviour in humans involve the use of neuroimaging Hariri, Drabant, Munoz, Kolachana, Mattay, Egan, et al. (2005); Hariri, Mattay, Tessitore, Kolachana, Fera, Goldman, et al. (2002), showed that variability in serotonin transporter (*SLC6A4*) genotype modulates amygdala response to fear faces. Using the same paradigm Meyer-Lindenberg, Kolachana, Gold, Olsh, Nicodemus, Mattay, et al. (2008) showed that polymorphisms in the arginine vasopressin receptor 1A (*AVPR1A*) gene (previously linked to autism) are related to the amygdala response to faces displaying fear or anger. Work from our and other groups has shown that variations in the cannabinoid receptor (*CNR1*) gene modulate striatal response to

happy faces (Chakrabarti & Baron-Cohen, 2006; Domschke, Dannlowski, Ohrmann, Lawford, Bauer, Kugel, et al., 2008). While the studies above rely on a more bottom-up response to emotion (since the task involves passive viewing of facial expressions, or doing a matching task), a recent imaging genetic study reported the genetic variation underlying cognitive component of empathy (Walter, Schnell, Erk, Arnold, Kirsch, Esslinger, et al., 2010). Future research should further characterize such endophenotypes in ASC in combination with ideal candidate genes. In this regard, a range of robust endophenotypes pertaining to autism and empathy have been put forward, both at the behavioural and neural levels (Lombardo, Baron-Cohen, Belmonte, & Chakrabarti, 2011; Losh, Adolphs, Poe, Couture, Penn, Baranek, et al., 2009).

The emerging picture that dimensional endophenotypes, rather than categorical diagnostic entities are useful targets for future genetic research is also reflected by the recent move to incorporate more dimensional measures in the new version of the DSM (DSM-5). This approach raises the issue of specificity, i.e. the endophenotypes may not be specific to certain categorical diagnostic entities. There is considerable evidence to suggest similarities in the social cognitive impairments between ASC and schizophrenia (Couture, Penn, Losh, Adolphs, Hurley, & Piven, 2010; King & Lord, 2011). Indeed, the polymorphism associated with differences in neural response in a ToM task was first reported from a GWA study of schizophrenia (Walter et al., 2010). The lack of disease-specificity of endophenotypic measures is mirrored by a similar overlap across diagnostic entities seen in genetic studies (Burbach & van der Zwaag, 2009; Guilmatre, Dubourg, Mosca, Legallic, Goldenberg, Drouin-Garraud, et al., 2009). The proposed future direction is therefore one where specific genetic loci will be characterized with their role in well-defined endophenotypes. One such genetic loci that has been well characterized is 7q11. Deletions in this locus are associated with Williams-Beuren Syndrome (where individuals are highly social), and duplications have been associated with ASC (Sanders, Hus, Luo, Murtha, Moreno-De-Luca, Chu, et al., 2011).

In closing, in this chapter we have presented a brief overview of genetics approaches to study empathy and autism. We have then discussed two recent genetic association experiments from our lab, one on autistic traits and empathy, and one on Asperger Syndrome. Finally, we have suggested potential avenues for future research, particularly using cross-validation through relevant endophenotypes. This combination of a functional hypothesis-driven search for candidate genes, alongside the development of fine-tuned quantitative phenotypic measures of brain and behaviour, will slowly bridge the gap between genes to cognition in the study of empathy and autism.

Acknowledgements

Parts of this chapter have been reprinted from (Chakrabarti and Baron-Cohen, 2011) and (Chakrabarti et al., 2009). We are grateful to Michael Lombardo, Ian Craig, Frank Dudbridge, and Lindsey Kent for valuable discussions.

References

- Abrahams, B. S., & Geschwind, D. H. (2008). Advances in autism genetics: on the threshold of a new neurobiology. *Nature Reviews Genetics* 9(5): 341–55.
- AGPC. (2007). Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genetics* 39: 319–28.
- APA. (1987). *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. Washington DC: American Psychiatric Association.
- Arking, D. E., Cutler, D. J., Brune, C. W., Teslovich, T. M., West, K., Ikeda, M., Rea, A., et al. (2008). A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *American Journal of Human Genetics* 82(1): 160–4.

- Ashley-Koch, A., Mei, H., Jaworski, J., Ma, D., Ritchie, M., Menold, M., et al. (2006). An analysis paradigm for investigating multi-locus effects in complex disease: Examination of three GABA A receptor subunit genes on 15q11-q13 as risk factors for autistic disorder. *Annals of Human Genetics* 70: 281–92.
- Auyeung, B., Baron-Cohen, S., Chapman, E., Knickmeyer, R., Taylor, K., & Hackett, G. (2006). Foetal testosterone and the child systemizing quotient. *European Journal of Endocrinology* 155(1): 123–30.
- Auyeung, B., Baron-Cohen, S., Ashwin, E., Knickmeyer, R., Taylor, K., Hackett, G., & Hines, M. (2009). Fetal testosterone predicts sexually differentiated childhood behavior in girls and in boys. *Psychological Science* 20(2): 144.
- Auyeung, B., Taylor, K., Hackett, G., & Baron-Cohen, S. (2010). Foetal testosterone and autistic traits in 18 to 24-month-old children. *Molecular Autism* 1: 11.
- Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simmonoff, E., Yuzda, E., & Rutter, M. (1995). Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological Medicine* 25: 63–77.
- Barnea-Goraly, N., Kwon, H., Menon, V., Eliez, S., Lotspeich, L., & Reiss, A. (2004). White matter structure in autism: preliminary evidence from diffusion tensor imaging. *Biological Psychiatry* 55(3): 323–6.
- Baron-Cohen, S. (2011). *Zero Degrees of Empathy: A New Theory of Human Cruelty*. London: Allen Lane.
- Baron-Cohen, S., Knickmeyer, R., & Belmonte, M. (2005). Sex Differences in the Brain: Implications for Explaining Autism. *Science* 310(5749): 819–23.
- Baron-Cohen, S., Ring, H., Bullmore, E., Wheelwright, S., Ashwin, C., & Williams, S. (2000). The amygdala theory of autism. *Neuroscience and Behavioural Reviews* 24: 355–64.
- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient (EQ). An investigation of adults with Asperger Syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders* 34: 163–75.
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism spectrum quotient (AQ): Evidence from Asperger Syndrome/high functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders* 31: 5–17.
- Baron-Cohen, S., & Hammer, J. (1997). Parents of children with Asperger syndrome: What is the cognitive phenotype? *Journal of Cognitive Neuroscience* 9(4): 548–54.
- Belmonte, M., Cook Jr, E., Anderson, G., Rubenstein, J., Greenough, W., Beckel-Mitchener, A., et al. (2004). Autism as a disorder of neural information processing: directions for research and targets for therapy. *Molecular Psychiatry* 9: 646–63.
- Bishop, D., Maybery, M., Maley, A., Wong, D., Hill, W., & Hallmayer, J. (2004). Using self-report to identify the broad phenotype in parents of children with autistic spectrum disorders: a study using the Autism-Spectrum Quotient. *Journal of Child Psychology and Psychiatry* 45(8): 1431–6.
- Bourgeron, T. (2009). A synaptic trek to autism. *Current Opinion in Neurobiology* 19(2): 231–4.
- Bowlby, J. (1969). *Attachment*. London: Hogarth Press.
- Burbach, J. P. H., & van der Zwaag, B. (2009). Contact in the genetics of autism and schizophrenia. *Trends in Neurosciences* 32(2): 69–72.
- Buxbaum, J., Silverman, J., Smith, C., Greenberg, D., Kilfarski, M., Reichart, J., Cook Jr, E., et al. (2002). Association between a GABRB3 polymorphism and autism. *Molecular Psychiatry* 7: 311–16.
- Campbell, D. B., Warren, D., Sutcliffe, J. S., Lee, E. B., & Levitt, P. (2010). Association of MET with social and communication phenotypes in individuals with autism spectrum disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 153(2): 438–46.
- Chakrabarti, B., & Baron-Cohen, S. (2011). Genes related to autistic traits and empathy. *From DNA to Social Cognition* 19–36.
- Chakrabarti, B., & Baron-Cohen, S. (2006). Empathizing: Neurocognitive developmental mechanisms and individual differences. *Progress in Brain Research*, 156(Special issue on “Understanding Emotions”), 406–17.

- Chakrabarti, B., Dudbridge, F., Kent, L., Wheelwright, S., Hill-Cawthorne, G., Allison, C., et al. (2009). Genes related to sex steroids, neural growth, and social-emotional behavior are associated with autistic traits, empathy, and Asperger syndrome. *Autism Research* 2(3): 157–77.
- Chakrabarti, B., Kent, L., Suckling, J., Bullmore, E., & Baron-Cohen, S. (2006). Variations in human cannabinoid receptor (*CNR1*) gene modulate striatal response to happy faces. *European Journal of Neuroscience* 23(7): 1944–8.
- Chapman, E., Baron-Cohen, S., Auyeung, B., Knickmeyer, R., Hackett, G., & Taylor, K. (2006). Foetal testosterone and empathy: evidence from the empathy quotient (EQ) and the “Reading the mind in the Eyes Test.” *Social Neuroscience* 1: 135–48.
- Collins, F., Guyer, M., & Chakravarti, A. (1997). Variations on a Theme: Cataloging Human DNA Sequence Variation. *Science* 278(5343): 1580–1.
- Conciatori, M., Stodgell, C., Hyman, S., O’Bara, M., Militeri, R., Bravaccio, C., et al. (2004). Association between the *HOXA1* A218G polymorphism and increased head circumference in patients with autism. *Biological Psychiatry* 55(4): 413–19.
- Constantino, J., Przybeck, T., Friesen, D., & Todd, R. (2000). Reciprocal social behaviour in children with and without pervasive developmental disorders. *Developmental and Behavioural Pediatrics* 1: 2–11.
- Constantino, J., & Todd, R. (2000). Genetic structure of reciprocal social behaviour. *American Journal of Psychiatry* 157, 2043–5.
- Constantino, J., & Todd, R. (2005). Intergenerational transmission of subthreshold autistic traits in the general population. *Biological Psychiatry* 57(6): 655–60.
- Corvin, A., Craddock, N., & Sullivan, P. F. (2010). Genome-wide association studies: a primer. *Psychological Medicine* 40: 1063–77.
- Courchesne, E., Campbell, K., & Solso, S. (2011). Brain growth across the life span in autism: age-specific changes in anatomical pathology. *Brain Research* 1380: 138–45.
- Courchesne, E., Pierce, K., Schumann, C. M., Redcay, E., Buckwalter, J. A., Kennedy, D. P., et al. (2007). Mapping early brain development in autism. *Neuron* 56(2): 399–413.
- Couture, S. M., Penn, D. L., Losh, M., Adolphs, R., Hurley, R., & Piven, J. (2010). Comparison of social cognitive functioning in schizophrenia and high functioning autism: more convergence than divergence. *Psychological Medicine* 40(04): 569–79.
- Crespi, B., & Badcock, C. (2008). Psychosis and autism as diametrical disorders of the social brain. *Behavioral and Brain Sciences* 31(03): 241–61.
- Da Silva, J. S., Hasegawa, T., Miyagi, T., Dotti, C. G., & Abad-Rodriguez, J. (2005). Asymmetric membrane ganglioside sialidase activity specifies axonal fate. *Nature neuroscience*, 8(5): 606–15.
- Davis, M., Luce, C., & Kraus, S. (1994). The heritability of characteristics associated with dispositional empathy. *Journal of Personality* 62(3): 369–91.
- Dawson, G., Carver, L., Meltzoff, A., Panagiotides, H., McPartland, J., & Webb, S. (2002). Neural correlates of face and object recognition in young children with autism spectrum disorder, developmental delay and typical development. *Child Development* 73: 700–17.
- DeLorey, T., Sahbaie, P., Hashemi, E., Homanics, G., & Clark, J. (2007). *Gabrb3* gene deficient mice exhibit impaired social and exploratory behaviors, deficits in non-selective attention and hypoplasia of cerebellar vermal lobules: A potential model of autism spectrum disorder. *Behavioural Brain Research* 187(2): 207–20.
- Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin improves “mind-reading” in humans. *Biological Psychiatry* 61(6): 731.
- Domschke, K., Dannlowski, U., Ohrmann, P., Lawford, B., Bauer, J., Kugel, H., et al. (2008). Cannabinoid receptor 1 (*CNR1*) gene: impact on antidepressant treatment response and emotion processing in major depression. *European Neuropsychopharmacology* 18(10): 751–9.
- Dudbridge, F. (2008). Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Human Heredity* 66(2): 87–98.

- Ebstein, R. P., Israel, S., Chew, S. H., Zhong, S., & Knafo, A. (2010). Genetics of Human Social Behavior. *Neuron* 65(6): 831–44.
- File, S. E., & Seth, P. (2003). A review of 25 years of the social interaction test. *European Journal of Pharmacology* 463(1–3), 35–53.
- Folstein, S., & Rutter, M. (1977). Infantile autism: A genetic study of 21 twin pairs. *Journal of Child Psychology and Psychiatry* 18: 297–321.
- Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., et al. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459(7246): 569–73.
- Good, C., Lawrence, K., Simon-Thomas, N., Price, C., Ashburner, J., Friston, K., et al. (2003). Dosage-sensitive X-linked locus influences the development of amygdala and orbitofrontal cortex, and fear recognition in humans. *Brain* 126, 1–16.
- Gottesman, I. I., & Gould, T. D. (2003). The endophenotype concept in psychiatry: etymology and strategic intentions. *American Journal of Psychiatry* 160(4): 636.
- Grimsby, J., Toth, M., Chen, K., Kumazawa, T., Klaidman, L., Adams, J., et al. (1997). Increased stress response and bold beta- phenylethylamine in MAOB- deficient mice. *Nature Genetics* 17: 206–10.
- Guilmatre, A., Dubourg, C., Mosca, A. L., Legallic, S., Goldenberg, A., Drouin-Garraud, V., et al. (2009). Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Archives of General Psychiatry* 66(9): 947–56.
- Happé, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience* 9(10): 1218–20.
- Hariri, A., Drabant, E., Munoz, K., Kolachana, B., Mattay, V., Egan, M., et al. (2005). A susceptibility gene for affective disorders and the response of the human amygdala. *Archives of General Psychiatry* 62(2): 146–52.
- Hariri, A., Mattay, V., Tessitore, A., Kolachana, B., Fera, F., Goldman, D., et al. (2002). Serotonin transporter genetic variation and the response of the human amygdala. *Science* 297: 400–3.
- Hoekstra, R., Bartels, M., Verweij, C., & Boomsma, D. (2007). Heritability of Autistic traits in the general population. *Archives of Pediatric and Adolescent Medicine* 161: 372–7.
- Hollander, E., Novotny, S., Hanratty, M., Yaffe, R., DeCaria, C., Aronowitz, B., & Mosovich, S. (2003). Oxytocin infusion reduces repetitive behaviors in adults with autistic and Asperger's disorders. *Neuropsychopharmacology* 28(1): 193–8.
- Holt, R., & Monaco, A. P. (2011). Links between genetics and pathophysiology in the autism spectrum disorders. *EMBO Molecular Medicine* 3(8): 438–50.
- Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child Development* 76(2): 356–70.
- ICD-10. (1994). *International Classification of Diseases*, Vol. 10. Geneva: World Health Organization.
- Ingudomnukul, E., Baron-Cohen, S., Wheelwright, S., & Knickmeyer, R. (2007). Elevated rates of testosterone-related disorders in women with autism spectrum conditions. *Hormones and Behavior* 51(5): 597–604.
- Insel, T., O'Brien, D., & Leckman, J. (1999). Oxytocin, vasopressin, and autism: is there a connection? *Biological Psychiatry* 45(2): 145–57.
- Jacob, S., Brune, C. W., Carter, C. S., Leventhal, B. L., Lord, C., & Cook Jr, E. H. (2007). Association of the oxytocin receptor gene (OXTR) in Caucasian children and adolescents with autism. *Neuroscience Letters* 417(1), 6–9.
- Jacquemont, M. L., Sanlaville, D., Redon, R., Raoul, O., Cormier-Daire, V., Lyonnet, S., et al. (2006). Array-based comparative genomic hybridization identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. *Journal of Medical Genetics* 43(11): 843–9.
- Jamain, S., Quach, H., Betancur, C., Råstam, M., Colineaux, C., Gillberg, I. C., et al. (2003). Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nature* 34, 27–9.

- Jones, A. P., Happé, F. G. E., Gilbert, F., Burnett, S., & Viding, E. (2010). Feeling, caring, knowing: different types of empathy deficit in boys with psychopathic tendencies and autism spectrum disorder. *Journal of Child Psychology and Psychiatry* 51(11): 1188–97.
- Just, M., Cherkassky, V., Keller, T., & Minshew, N. (2004). Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. *Brain* 127: 1811–21.
- Keller, T. A., Kana, R. K., & Just, M. A. (2007). A developmental study of the structural integrity of white matter in autism. *Neuroreport*, 18(1): 23–28.
- King, B. H., & Lord, C. (2011). Is schizophrenia on the autism spectrum? *Brain research*, 1380, 34–41.
- Knafo, A., Zahn-Waxler, C., Van Hulle, C., Robinson, J. L., & Rhee, S. H. (2008). The developmental origins of a disposition toward empathy: Genetic and environmental contributions. *Emotion*, 8(6):737–52.
- Knickmeyer, R., Baron-Cohen, S., Fane, B. A., Wheelwright, S., Mathews, G. A., Conway, G. S., et al. (2006). Androgens and autistic traits: A study of individuals with congenital adrenal hyperplasia. *Hormones and Behavior* 50(1): 148–53.
- Knickmeyer, R., Baron-Cohen, S., Raggatt, P., & Taylor, K. (2005). Foetal testosterone, social cognition, and restricted interests in children. *Journal of Child Psychology and Psychiatry* 46(2): 198–210.
- Koido, K., Köks, S., Nikopensus, T., Maron, E., Altmäe, S., Heinaste, E., et al. (2004). Polymorphisms in wolframin (WFS1) gene are possibly related to increased risk for mood disorders. *International Journal of Neuropsychopharmacology* 8(2): 235–44.
- Koks, S., Planken, A., Luuk, H., & Vasar, E. (2002). Cat odour exposure increases the expression of wolframin gene in the amygdaloid area of rat. *Neuroscience Letters* 322(2): 116–20.
- Kosfeld, M., Heinrichs, M., Zak, P., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature* 435(2): 673–6.
- Kudwa, A. E., Bodo, C., Gustafsson, J., & Rissman, E. F. (2005). A previously uncharacterized role for estrogen receptor: Defeminization of male brain and behavior. *Proceedings of the National Academy of Sciences, USA* 102(12): 4608.
- Kuribayashi, I., Nomoto, S., Massa, G., Oostdijk, W., Wit, J., Wolffenbuttel, B., et al. (2005). Steroid 11-beta-hydroxylase deficiency caused by compound heterozygosity for a novel mutation, p G314R, in one CYP11B1 allele, and a chimeric CYP11B2/CYP11B1 in the other allele. *Hormone Research* 63: 284–93.
- LaSalle, J. M., & Yasui, D. H. (2009). Evolving role of MeCP2 in Rett syndrome and autism. *Epigenomics* 1(1): 119–30.
- Loehlin, J., & Nichols, R. (1976). *Heredity, Environment and Personality*. Austin: University of Texas Press.
- Lombardo, M. V., Baron-Cohen, S., Belmonte, M. K., & Chakrabarti, B. (2011). Neural endophenotypes for social behaviour in autism spectrum conditions. In: J. Decety & J. T. Cacioppo (Eds), *Oxford Handbook of Social Neuroscience*. Oxford: Oxford University Press.
- Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., et al. (1989). Autism diagnostic observation schedule: A standard observation of communicative and social behaviour. *Journal of Autism and Developmental Disorders* 19: 185–212.
- Losh, M., & Piven, J. (2007). Social-cognition and the broad autism phenotype: identifying genetically meaningful phenotypes. *Journal of Child Psychology and Psychiatry* 48(1): 105–12.
- Losh, M., Sullivan, P. F., Trembath, D., & Piven, J. (2008). Current developments in the genetics of autism: from phenome to genome. *Journal of Neuropathology and Experimental Neurology* 67(9): 829–37.
- Losh, M., Adolphs, R., Poe, M. D., Couture, S., Penn, D., Baranek, G. T., et al. (2009). Neuropsychological profile of autism and the broad autism phenotype. *Archives of General Psychiatry* 66(5): 518–26.
- Lutchmaya, S., Baron-Cohen, S., & Raggatt, P. (2002). Foetal testosterone and vocabulary size in 18- and 24-month-old infants. *Infant Behavior and Development* 24(4): 418–24.
- Maltepe, E., Keith, B., Arsham, A., Brorson, J., & Simon, M. (2000). The role of ARNT2 in tumor angiogenesis and the neural response to hypoxia. *Biochemistry and Biophysics Research Communication* 273(1): 231–8.

- Matthews, K., Batson, C., Horn, J., & Rosenman, R. (1981). "Principles in his nature which interest him in the fortune of others ...": The heritability of empathic concern for others. *Journal of Personality* 49(3): 237–47.
- Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality* 40: 525–43.
- Meyer-Lindenberg, A., Kolachana, B., Gold, B., Olsh, A., Nicodemus, K. K., Mattay, V., et al. (2008). Genetic variants in AVPR1A linked to autism predict amygdala activation and personality traits in healthy humans. *Molecular Psychiatry* 14(10): 968–75.
- Michaud, J., DeRossi, C., May, N., Holdener, B., & Fan, C. (2000). ARNT2 acts as the dimerization partner of SIM1 for the development of the hypothalamus. *Mechanics of Development* 90(2): 253–61.
- Minshew, N. J., & Williams, D. L. (2007). The new neurobiology of autism: Cortex, connectivity, and neuronal organization. *Archives of Neurology* 64(7): 945.
- Muscarella, L., Guarnieri, V., Sacco, R., Militeri, R., Bravaccio, C., Trillo, S., et al. (2007). *Hoxa1* gene variants influence head growth rates in humans. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 144B(3): 388–90.
- Nagarajan, R. P., Patzel, K. A., Martin, M., Yasui, D. H., Swanberg, S. E., Hertz-Picciotto, I., et al. (2008). MECP2 promoter methylation and X chromosome inactivation in autism. *Autism Research* 1(3): 169–78.
- Park, J. M., Lee, E. J., Ramakrishna, S., Cha, D. H., & Baek, K. H. (2008). Association study for single nucleotide polymorphisms in the *CYP17A1* gene and polycystic ovary syndrome. *International Journal of Molecular Medicine* 22(2): 249.
- Pharoah, P. D., Tyrer, J., Dunning, A. M., Easton, D. F., & Ponder, B. A. (2007). Association between common variation in 120 candidate genes and breast cancer risk. *PLoS Genetics* 3(3): e42.
- Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 466(7304): 368–72.
- Piven, J., Palmer, P., Jacobi, D., Childress, D., & Arndt, S. (1997). Broader autism phenotype: evidence from a family history study of multiple-incidence autism families. *American Journal of Psychiatry* 154: 185–90.
- Preston, S. D., & De Waal, F. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences* 25(1): 1–20.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81(3): 559–75.
- Robinson, E. B., Koenen, K. C., McCormick, M. C., Munir, K., Hallett, V., Happe, F., et al. (2011). Evidence that autistic traits show the same etiology in the general population and at the quantitative extremes (5%, 2.5%, and 1%). *Archives of General Psychiatry* 68(11): 1113–21.
- Rogers, J., Viding, E., Blair, R. J., Frith, U., & Happe, F. (2006). Autism spectrum disorder and psychopathy: shared cognitive underpinnings or double hit? *Psychological Medicine* 36(12): 1789–8.
- Rushton, J., Fulker, D., Neale, M., Nias, D., & Eysenck, H. (1986). Altruism and aggression: The heritability of individual differences. *Journal of Personality and Social Psychology* 50(6): 1192–8.
- Ruta, L., Ingudomnukul, E., Taylor, K., Chakrabarti, B., & Baron-Cohen, S. (2011). Increased serum androstenedione in adults with autism spectrum conditions. *Psychoneuroendocrinology* 36: 1154–63.
- Sahyoun, C. P., Belliveau, J. W., & Mody, M. (2010). White matter integrity and pictorial reasoning in high-functioning children with autism. *Brain and Cognition* 73(3): 180–8.
- Sanders, S. J., Hus, V., Luo, R., Murtha, M. T., Moreno-De-Luca, D., Chu, S. H., et al. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 70(5): 863–85.
- Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., et al. (2007). Strong association of *de novo* copy number mutations with autism. *Science* 316(5823): 445–9.

- Shih, P., Shen, M., Öttl, B., Keehn, B., Gaffrey, M. S., & Müller, R. A. (2010). Atypical network connectivity for imitation in autism spectrum disorder. *Neuropsychologia* 48(10): 2931–9.
- Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences*, 1156(The Year in Cognitive Neuroscience 2009), 81–96.
- Sofroniew, M., Howe, C., & Mobley, W. (2001). Nerve growth factor signaling, neuroprotection, and neural repair. *Annual Review of Neuroscience* 24: 1217–81.
- Sundaram, S. K., Kumar, A., Makki, M. I., Behen, M. E., Chugani, H. T., & Chugani, D. C. (2008). Diffusion tensor imaging of frontal lobe in autism spectrum disorder. *Cerebral Cortex* 18: 2659–65.
- Sung, Y., Dawson, G., Munson, J., Estes, A., Schellenberg, G., & Wijsman, E. (2005). Genetic investigation of quantitative traits related to autism: Use of multivariate polygenic models with ascertainment adjustment. *American Journal of Human Genetics* 76(1): 68–81.
- Tops, M., Van IJzendoorn, M. H., Riem, M. M., Boksem, M. A., & Bakermans-Kranenburg, M. J. (2011). Oxytocin receptor gene associated with the efficiency of social auditory processing. *Frontiers in Psychiatry*, 2.
- Villalobos, M. E., Mizuno, A., Dahl, B. C., Kemmotsu, N., & Müller, R. A. (2005). Reduced functional connectivity between V1 and inferior frontal cortex associated with visuomotor performance in autism. *NeuroImage* 25(3): 916–25.
- Wagner, G. C., Reuhl, K. R., Cheh, M., McRae, P., & Halladay, A. K. (2006). A new neurobehavioral model of autism in mice: pre- and postnatal exposure to sodium valproate. *Journal of Autism and Developmental Disorders* 36(6): 779–93.
- Walter, H., Schnell, K., Erk, S., Arnold, C., Kirsch, P., Esslinger, C., et al. (2010). Effects of a genome-wide supported psychosis risk variant on neural activation during a theory-of-mind task. *Molecular Psychiatry* 16(4): 462–70.
- Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J. T., Abrahams, B. S., et al. (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 459(7246): 528–33.
- Wass, S. (2011). Distortions and disconnections: disrupted brain connectivity in autism. *Brain and Cognition* 75(1): 18.
- Welchew, D. (2005). Functional dysconnectivity of the medial temporal lobe in Asperger's syndrome. *Biological Psychiatry* 57: 991–8.
- Wermter, A. K., Kamp-Becker, I., Hesse, P., Schulte-Körne, G., Strauch, K., & Remschmidt, H. (2009). Evidence for the involvement of genetic variation in the oxytocin receptor gene (*OXTR*) in the etiology of autistic disorders on high-functioning level. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics* 153B(2): 629–639.
- Wheelwright, S., Auyeung, B., Allison, C., & Baron-Cohen, S. (2010). Defining the broader, medium and narrow autism phenotype among parents using the autism spectrum quotient (AQ). *Molecular Autism* 1: 10.
- Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., ... & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences* 107(11): 5238–41.
- Wu, N., Li, Z., & Su, Y. (2012). The association between oxytocin receptor gene polymorphism (*OXTR*) and trait empathy. *Journal of affective disorders* 138(3): 468–72.
- Wu, S., Jia, M., Ruan, Y., Liu, J., Guo, Y., Shuang, M., et al. (2005). Positive association of the oxytocin receptor gene (*OXTR*) with autism in the Chinese Han population. *Biological Psychiatry* 58(1): 74–7.
- Ylisaukko-oja, T., Nieminen-von Wendt, T., Kempas, E., Sarenius, S., Varilo, T., von Wendt, L., et al. (2004). Genome-wide scan for loci of Asperger syndrome. *Molecular Psychiatry* 9(2): 161–8.
- Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of concern for others. *Developmental Psychology* 28: 126–36.

Section 3

Psychiatric, neurodevelopmental, and neurological disorders

This page intentionally left blank

Theory of mind in deaf children: Illuminating the relative roles of language and executive functioning in the development of social cognition

Jennie Pyers and Peter A. de Villiers

In the mid-1990s several independent programs of research and theorizing proposed that studies of deaf children could illuminate the role of language in children's theory of mind (ToM) or more broadly in their social cognitive development. For example, in a theoretical paper about the relationship between language and thought, Jackendoff (1996) argued that language was necessary for making explicit judgments about the truth and falsity of propositions. In a footnote he noted that a colleague had recognized that an implication of his theory was that language-delayed deaf children would have difficulty with false-belief (FB) tasks. Just at that time, two independent research groups confirmed that hypothesis: one with late signing deaf children (Peterson & Siegal, 1995) and the other with orally-taught deaf children (Gale, de Villiers, de Villiers, & Pyers, 1996). Indeed, deaf children with hearing parents provide a strong test of the hypothesis that language plays a causal role in ToM development because these children experience varying degrees of language delay, but typically have a normal IQ and active sociability.

In this chapter we first discuss some crucial methodological challenges in studying ToM development in deaf children. In the light of these methodological issues, we review the most comprehensive studies of deaf children's explicit FB reasoning and how language and executive function (EF) ability does and does not affect this development. Since the initial ground-breaking research, the picture of deaf children's ToM has gotten considerably more complex. The resulting picture confirms other arguments that many different components constitute a fully articulated "theory of mind" (e.g. Wellman & Liu, 2004), and each of these components may be differentially affected by language acquisition and/or executive functioning. We will show that some social cognitive understandings, such as those embodied in deceptive games with low verbal requirements, do not appear to be delayed by language-impairment resulting from deafness and are predicted by inhibitory control (an aspect of EF) not language skills. Others, such as reasoning about states of knowledge and ignorance and explicit judgments about false beliefs are considerably delayed and closely predicted by language, but not by deaf children's EF.

Methodological issues

The existing literature points to significant limitations in FB reasoning in deaf children, but different studies report widely variable ages at which deaf children of hearing parents succeed on either high-verbal or low-verbal measures of FB understanding. Some of this variability may be

attributed to the diversity of the deaf samples studied, and some may be attributed to methodological differences in the assessment of deaf children.

Key to what makes deaf children an ideal population in which to investigate the relative contributions of language and EF to ToM development is the diversity of deaf children's language experience. Typically two distinct populations of deaf children are included in studies of ToM. Deaf children born to deaf parents (DoD) function as a control population because, despite their deafness, they have native exposure to an accessible first language, a natural sign language. Thus, they have normal language acquisition, albeit in a different modality from typically-hearing (TH) children. On the other hand, deaf children born to hearing parents (DoH) have greater variability in their language experience: some acquire a natural sign language, some learn only a manually coded version of the spoken language of their community (e.g. signed English), and others never learn to sign and are exposed only to oral language. Whatever their language experience, DoH children typically display some degree of language delay without any corresponding congenital cognitive deficit (Marschark, 1993). While the language delay makes DoH children the ideal population in which to test the effects of language on ToM, the diversity within this group can yield widely different levels of performance. As such, studies that provide the strongest information about the effects of language on ToM in deaf children include detailed information about the children's language experience as well as measures of their ability in their preferred language, signed or spoken.

Background variables, beyond language experience, can also impact children's performance. Relative to the TH children who are commonly recruited from university-affiliated preschools, deaf children in the United States come from more diverse socio-economic backgrounds (SES) and may have other physical and cognitive challenges. Several of these background variables that might impact children's ToM performance are summarized in Table 19.1. Ideally, researchers should include large samples of deaf children in their studies to minimize the effect of these other variables. However, with only 0.64% of children in the United States diagnosed as hard-of-hearing or deaf (Mitchell, 2006), recruiting large samples of deaf children remains difficult and expensive. Alternatively, researchers could address the variability in deaf samples either by finding a TH sample that closely matches the deaf sample on SES and on general cognitive ability, or by statistically controlling for all of the background variables, a difficult task with the usual small sample sizes.

Approaches to deaf education also shape the characteristics of the deaf population to which researchers have the easiest access. The current educational practice in the United States and in many other Western countries is to work toward deaf children's full integration into educational programs for typically-developing children. As soon as deaf children develop the social, cognitive, and linguistic foundations deemed necessary to succeed, they are often placed in mainstream schools with TH children, and they learn with the support of sign language interpreters and/or additional amplification services. Deaf children can enter the mainstream in some cases as early as preschool or as late as high school. Some deaf children never transfer into integrated programs for a variety of reasons including parental choice or not having acquired the foundational skills that the program considers necessary for educational success. The practice of transferring deaf children to integrated programs means that many older deaf children with age-appropriate language and cognitive abilities are no longer enrolled in programs that educate only deaf children. Thus studies that target only deaf children enrolled in special schools or programs for the deaf may not be representative of the most successful deaf children. This limitation is greater for studies that include children beyond the preschool and kindergarten years; deaf children start to move into the mainstream during the early elementary years, leaving behind classrooms where children who

Table 19.1 Background variables that must be accounted for to best interpret data from studies of deaf children

SES	SES affects FB performance in typically developing children, with children from higher SES families outperforming children from less privileged backgrounds (Cicchetti, Rogosch, Maughan, Toth, & Bruce, 2003; Shatz, Diesendruck, Martinez-Beck, & Akar, 2003). When deaf and hearing children are matched for SES and non-verbal IQ, native signers perform equivalently or even better than TH children on FB tasks, and the reported difference between DoH and DoD children is much smaller (Courtin, 2000; Schick, de Villiers, de Villiers, & Hoffmeister, 2007).
Non-verbal intelligence	The inclusion of measures of non-verbal intelligence such as spatial working memory ensures that the children in the sample have a normal intellectual level and reduces the likelihood that any limitations observed in ToM performance are due to limitations in general intellectual ability.
Physical or cognitive impairments aside from deafness	Deafness sometimes is co-morbid with other physical or intellectual disabilities that may affect performance on FB tasks independently of language (Marschark, 1993). Such information gathered from the school helps guide researchers as to whether it is necessary to exclude these children from the final analyses.
General language ability	For deaf children exposed only to a spoken language, measures of spoken language development that have been validated with deaf children are appropriate. In the United States, such measures include the CELF and the Rhode Island Test of Language Structure (Engen & Engen, 1983). The assessment of sign language ability is much more difficult without standardized measures. Several researchers have developed their own language measures to compare sign language development in DoD and DoH children (e.g. Schick et al., 2007), but these measures do not always allow for cross-linguistic comparisons of signed and spoken language acquisition. Signed translations of spoken language measures are not valid measures of natural sign language acquisition.
Age of first language exposure/age of amplification	Age of first exposure to language impacts language ability (Mayberry & Lock, 2003). For non-native signing children, age of first language exposure is typically the age at which they entered a signing program; for oral deaf children it is usually the age when they receive their first amplification to enhance their auditory access to spoken language. In the United States, all children who are identified with a hearing-loss at birth receive state-supported early intervention that provides both auditory and language intervention. However, some families first choose an oral-only methodology, but later introduce their child to a sign language when the child struggles with acquiring a spoken language.
Language exposure in the home	For signing children, the degree to which hearing parents learn a sign language varies. Some parents become more fluent than others (Moeller & Schick, 2006; Vaccari & Marschark, 1997). In addition, some hearing parents of deaf children speak more than one spoken language in the home, reducing the amount of English heard by orally taught deaf children.

have other needs beyond deafness outnumber children who otherwise exhibit typical cognitive, physical, and social development.

Beyond careful consideration of the variability in deaf children's backgrounds, researchers' testing methods can impact the children's performance. First, most deaf children primarily acquire linguistic information through the visual channel by viewing signs or reading lips. Some traditional FB measures engage TH children's ability to monitor visual and auditory information simultaneously: a TH child can readily look away from the experimenter to view action taking place in a dollhouse and still listen to the experimenter's narrative. For deaf children the traditional tasks often require shifting their visual attention from the tester to the test stimuli. Modified ToM measures can reduce the demands made on deaf children's visual attention. For example, some researchers have designed storybooks that can be propped up underneath the experimenter's face and in front of which the experimenter can sign (Schick et al., 2007), an adaptation that many signing parents use with their deaf children (Lieberman, Hatrak, & Mayberry, 2011).

Many oral deaf children rely not only on lip-reading, but also on what auditory information they can glean using their amplification systems. In the educational setting, teachers use a variety of amplification systems that transmit their voices directly to the hearing aids of the children in their classrooms, filtering out ambient background noise. Experimenters should also use such equipment when working with deaf children with some usable audition to maximize spoken language access. Close work with educators and speech therapists can create a testing situation that elicits an oral deaf child's best performance.

For signing children, deaf native signing experimenters elicit children's best linguistic performance. When a deaf native signing experimenter is not available, deaf or hearing fluent signers can also serve as experimenters. Some evidence indicates that deaf signers perform differently in the presence of non-fluent hearing signers, modifying their signs to conform less to the grammatical rules of the sign language (Cokely, 1983; Lucas & Valli, 1989). The least optimal testing situation is one in which a hearing non-signer administers the test and the language of the experimenter is translated by a sign language interpreter, commonly the classroom interpreter used in the program. In this situation, the child has to shift attention from the experimenter, to the interpreter, to the test stimuli. We know little about how well signing deaf children can shift their visual attention in such a situation (Corina & Singleton, 2009), and we know even less about how effectively young deaf children can use an interpreter. In studies with educational interpreters in the United States, 60% of the interpreters did not demonstrate advanced enough skills to ensure full access for children (Schick, Williams, & Kupermintz, 2006), and they were particularly deficient in conveying affect and prosody (Schick, 2004). More disturbingly, educational interpreters who are assigned to work with young preschool and elementary aged children typically scored the weakest on an assessment of educational interpretation (Schick et al., 2006). The very real limitations of using an interpreter to translate a non-signing experimenter's instructions compromises the researcher's ability to compare deaf children's performance to that of TH children who are tested without the intervention of an interpreter. Low ToM performance in an interpreted situation may have more to do with task translation issues and controlling visual attention than true limitations in ToM reasoning.

Finally, to answer the question of whether language experience impacts ToM, researchers must work to ensure that any struggles observed in language-delayed deaf children are specifically associated with their general language ability, not just their ability to follow the instructions of the task. One way researchers have attempted to address this issue is to develop non-verbal or minimally verbal tasks that do not require children to understand complex language to succeed. Some

minimally verbal measures are classic FB tasks presented as picture sequences that are administered with minimal language support and that require children to make a forced choice between two or more alternatives (de Villiers & de Villiers, 2012; Pyers & Senghas, 2009, Schick et al., 2007; Woolfe, Want, & Siegel, 2002). Other minimally verbal tasks have been adapted from behavioral tasks used with non-human primates. These tasks engage the children in hide-and-seek games where they have to monitor the knowledge state of an informed or an uninformed confederate (Figuera-Costa & Harris, 2001; Schick et al., 2007).

A comprehensive study of false belief reasoning

Schick et al. (2007) most closely approximates the ideal design and control considerations laid out above. Unlike most of the previous work, Schick et al. tested a substantial sample of both orally taught ($n = 86$) and ASL signing ($n = 90$) deaf children to provide sufficient statistical power. Most importantly, sufficient numbers of native-signing DoD 4-year-olds were studied so that their performance on a variety of low and high verbal ToM tasks could be compared with that of TH children matched on SES, age, and non-verbal IQ. In addition extensive assessment of the children's language acquisition was carried out, including measures of expressive and receptive vocabulary; receptive general syntax; and comprehension of tensed false complement clauses with verbs of communication (following de Villiers & Pyers, 2002). The oral deaf children were assessed in spoken English; the signing children (both the DoD and DoH groups) in American Sign Language (ASL). Special tests were developed to assess the ASL vocabulary, general syntax, and complement comprehension of the signing children. All of the testing was carried out by examiners with appropriate qualifications for working with either signing or oral deaf children: the signing participants by deaf examiners with native-signing ASL skills; and the oral deaf children by testers familiar with the speech of deaf children.

Schick et al. (2007) report several primary findings. First, language-delayed deaf children with hearing parents, whether they were educated in oral or signing schools, performed significantly worse on both high and low-verbal FB tasks than both the native-signing deaf children and the TH controls, indicating that the language demands of the standard tasks were not the cause of their poorer performance. Second, deafness alone does not affect ToM performance because native-signing deaf children and the TH control children performed at an equivalent level to each other on both the standard verbal FB tasks and the low-verbal analogs of the ToM tasks.

Hierarchical linear regression analyses determined whether background variables (age, hearing loss, non-verbal IQ, and sequence memory) or various aspects of the children's language skills were predictive of their reasoning about states of knowledge and FB. Importantly, the same pattern of predictors was found for the oral deaf children (tested in spoken English) and the signing deaf children (tested in ASL). The performance of both the oral and the signing deaf children on the standard verbal FB tasks was independently predicted by age, receptive vocabulary, and the children's processing of tensed false complement clauses with verbs of communication (as de Villiers & Pyers (2002) found for hearing preschoolers). Levels of hearing loss, non-verbal IQ, and sequence memory were not independent predictors of FB reasoning once age was controlled for. And comprehension of general syntactic features of spoken English or ASL was not as predictive of FB reasoning as vocabulary or processing of complement clauses. Passing the low verbal ToM tasks was independently predicted by age and processing of false complement clauses with communication verbs. The other background and language measures were not significant independent predictors for either oral or signing deaf children.

Other studies of false belief reasoning in deaf children

Two other studies using low verbal FB reasoning tasks with native signing and late signing deaf children add important nuances to the findings of Schick et al. (2007), although they tested older samples of deaf children and did not provide the same degree of age and SES matching of deaf children with a control group of TH children. Woolfe et al. (2002) found that native signing children performed significantly better than late signers on a low verbal “thought bubble” test of FB reasoning, even though the two groups of children did not differ in their raw scores on a standardized test of British Sign Language comprehension. They argue that the native signing children’s early exposure to comprehensible conversation and language about their own and others’ mental states therefore is a more important factor in their ToM reasoning than their current knowledge of sign language syntax. Note, however, that the BSL assessment administered by Woolfe et al did not assess the children’s comprehension of false complement clauses. Schick et al. (2007) also found that general ASL syntax was not a significant predictor of ToM in signing deaf children; processing of false complements with verbs of communication was the strongest predictor.

Two studies by Meristo, Falkman, Hjelmquist, Tedoldi, Surian, & Siegal (2007) support the importance of ongoing comprehensible language and conversation in facilitating development of a well-articulated ToM. Native signers of Italian Sign Language (ISL) who were educated in bilingual educational environments that used ISL as well as spoken language outperformed native signers educated in oral-only instructional environments on the thought bubble FB reasoning task devised by Woolfe et al. (2002). There were no significant differences between native signers in oral environments, late signers in oral environments, and late signers in bilingual-bicultural schools. As was the case in the study by Woolfe et al., the ISL native signers in the two different educational environments performed at the same level on a test of their comprehension of ISL (an ISL translation of the test of comprehension of BSL). A second study of Estonian and Swedish deaf children showed similar findings when the children were tested on a battery of verbal FB and advanced ToM tasks. Again the authors stress the crucial role of fluent, comprehensible communication throughout the day for optimizing the ToM development of deaf children, even when the children have input in a natural sign language at home.

Finally, research on Nicaraguan deaf signers by Pyers & Senghas (Pyers, 2005; Pyers & Senghas, 2009) supports the notion that adult deaf individuals may acquire elaborate social interactional skills and function well in their communities, but unless they have acquired an appropriately complex syntax in their sign language, they may still fail low-verbal, but explicit FB reasoning tasks. Adults who were members of the early cohorts in the evolution of Nicaraguan Sign Language (NSL) were much poorer at reasoning about characters’ FB (or emotions based on FB) than younger individuals from the later sign language cohorts. The older, early cohort adults also had less complex and formally elaborated syntax in their NSL. In addition, when called upon to describe what was happening in brief videotaped scenarios of mistakes and deceptive actions (see de Villiers and Pyers, 2001), the older cohort used more language involving physical causation and desire-based explanations, while the younger cohort produced more mental state explanations with belief and knowledge verbs and complement clauses (Pyers, 2005).

All of these studies argue that language plays a crucial role in the ToM development of deaf children (and by extension, of TH children as well). Delays in the acquisition of mental state vocabulary, complex aspects of syntax that may enable representation of the content of false beliefs (de Villiers & de Villiers, 2009), and impoverished and less comprehensible communication may impair deaf children’s understanding of cognitive states, especially in situations where they are false. However, could the mechanism by which language delay has its impact on ToM understanding in deaf

children not be directly on the underlying conceptual development, but on the executive functioning skills that are needed for the explicit reasoning tasks by which the children are assessed?

Executive functioning and false belief reasoning

An influential account of children's mastery of explicit reasoning about FB appeals to the development or maturation of executive functioning (EF) skills between the ages of 3 and 5. EF skills include the planning, monitoring, and control of behavior. Three aspects of the EF system have received most of the attention in research and theory about ToM development. First, inhibitory control, especially in situations of conflict between competing tasks or alternative behaviors (Carlson & Moses, 2001; Carlson, Moses, & Breton, 2002). Secondly, flexible rule following where there are conditional rules or set shifting when there are well-learned competing rules that the child needs to ignore (Frye, Zelazo, & Palfrai, 1995). Thirdly, working memory, which enables the child to keep the competing alternatives in mind (Davis & Pratt, 1995; Keenan, 2000).

Logical analysis of the usual FB reasoning tasks suggests that succeeding at them requires each of these three skills. In unseen location change tasks the child has to remember where the desired object was before and where it has been moved to and resist the lure of the current location of the object (reality) in order to respond correctly in terms of the false belief of the relevant character. Similarly, in the unexpected contents task the child is explicitly asked to remember what they thought was in the box before they looked inside, and therefore, they must inhibit the tendency to respond with what they have now seen has been substituted for the box's usual contents. Several studies of TH children have reported significant correlations between one or more of these features of EF and ToM development, even when age and verbal IQ are controlled for (Carlson & Moses, 2001; Carlson, Moses, & Breton, 2002; Frye et al., 1995; Davis & Pratt, 1995).

EF and deafness

Several studies investigating EF and ToM in deaf children stand in contrast to the findings with TH children, and report that the inhibitory control and set shifting skills of deaf children do not seem to be closely related to their explicit reasoning about FB in verbal or low-verbal tasks. Woolfe et al. (2002) tested age-matched native signers and late signers in a version of the Wisconsin Dimensional Card Sort and found that although the native signers were significantly better than the late signers on a low-verbal FB reasoning task, there was no difference between the two groups on the set-shifting task. Similarly, Meristo & Hjelmquist (2009) tested three groups of deaf students matched for age and non-verbal IQ: bilingually-instructed native signers, orally-instructed native signers, and bilingually-instructed late signers. The children were given a battery of verbal ToM and executive function tasks (administered in sign language by native signing research assistants). Although the native-signing deaf children from the bilingual educational settings were significantly better than the other deaf groups on the ToM reasoning tasks, there were no significant differences between the groups in verbal working memory (backwards digit span), set shifting (on the Wisconsin Card Sorting test) or conflict inhibitory control. Furthermore, when age and non-verbal IQ were partialled out, only verbal working memory was significantly correlated with FB reasoning on the standard verbal tasks. There was no correlation between conflict inhibition or set shifting and FB reasoning for the deaf children.

de Villiers & de Villiers (2012) studied 45 oral deaf children and 45 TH controls on a battery of EF, language, deception, and FB reasoning tasks. The younger of the deaf children (average age 5;3, range 4–6, $n = 29$) were closely matched with 18 of the hearing children (average age

5;2, range 4–6) in age and non-verbal sequence memory. There were no significant differences between these two matched groups of children on widely used EF tasks that included a two measures of conflict inhibitory control (the Day-Night Stroop test and the Knock-Tap hand game) or on a dimensional card sort set shifting task. However, the deaf children were significantly worse than the hearing children on both the standard verbal FB unseen object displacement and unexpected contents tasks and on two low verbal “thought bubble” tests of FB understanding based on the procedures used by Woolfe et al. (2002) and Schick et al. (2007). Furthermore, the deaf children were significantly impaired relative to the TH children in their mastery of general English sentence syntax and their memory for false complement clauses with verbs of communication (de Villiers & Pyers, 2002; Schick et al., 2007). For both the deaf and the TH groups, performance on the FB tasks was independently predicted by the children’s language and especially by their processing of false complement clauses, even when age and sequence memory were controlled for. None of the EF measures predicted FB reasoning in the deaf children, and they were weaker predictors than the language measures were for the hearing children.

Taken together, these studies with deaf children show that while EF may be necessary for success at explicit FB reasoning tasks, these skills are not the proximal predictors of deaf children’s level of performance on those tasks (see also de Villiers, 2005). The strongest predictors seem to be the children’s language skills.

Interestingly, on two low verbal deception games (the sticker-in-the-hand game and a deceptive pointing game) there were no significant differences between the deaf children and the hearing children in level of performance (de Villiers & de Villiers, 2012). The children’s deception scores were significantly predicted by their sequence memory and their inhibitory control, not by their language. Thus, there seems to be a dissociation of deception and explicit FB reasoning tasks for deaf children: the deaf children were on a par with their hearing peers on deception games, but showed significant delays in explicit FB reasoning even when the language demands of the tasks were minimized. de Villiers & de Villiers (2012) suggested that deception at this level could be handled by behavior rules without explicit representation of mental states (cf. Perner, 2010; Poivinelli & Vonk, 2004; Ruffman, Taumoepeau, & Perkins, 2012).

Continued growth in ToM reasoning in deaf individuals

A few of the earlier cross-sectional studies of FB reasoning in deaf individuals suggested that their impaired language acquisition and impoverished communication might produce lasting deficits in ToM, hinting at a critical period for the development of a conceptual understanding of FB (Edmondson, 2006; Morgan & Kegl, 2006; Russell, Hosie, Gray, Scott, Hunter, Banks, et al., 1998). However, two more complete longitudinal studies have documented acquisition of FB understanding even as late as early adulthood. Deaf Australian elementary school children showed delayed, albeit eventual, acquisition of FB understanding on standard verbal tasks (Wellman, Fang, & Peterson, 2011). Some members of a population of adult Nicaraguan signers who, as children, acquired a new, emerging sign language that had limited mental-state vocabulary initially failed low-verbal FB tasks, but after the very same signers acquired verbs of belief and knowledge as adults, they subsequently improved their FB performance on a low verbal FB task (Pyers & Senghas, 2009). This pattern of late acquisition provides strong evidence that the acquisition of FB understanding is not bounded by a critical period. Crucially, all participants in the study of Nicaraguan signers had otherwise typical social and environmental experience—their sole limitation was related to the complexity of their mental-state language. In extreme cases of deprivation—nutritional, social, and language—late acquisition of FB understanding may not be possible.

Development of other aspects of theory of mind

As in the case of research on TH children, study of ToM development in deaf individuals has been dominated by research on FB understanding in preschool and early elementary school. However, ToM development begins in infancy as children begin to exhibit attention to other's minds by following the eye-gaze of others (Brooks & Meltzoff, 2002), engaging in joint attention (Tomasello, Carpenter, Call, Behne, & Moll, 2005), and understanding others' goals and intentions (Meltzoff, 1995). Similarly, FB understanding in the late preschool years is preceded by an understanding of how knowledge is acquired, specifically how seeing and knowing are related (Flavell, 1992; Pratt & Bryant, 1990), and of non-belief mental states, such as desires (Bartsch & Wellman, 1989). This section of the chapter reviews the research on those other components of ToM development in deaf children and considers the way in which language delay may impact them.

Intention and desires

Children's first insight into others' minds emerges when they first see humans as intentional agents. Between 10–12 months of age, infants show sensitivity in looking-time measures to the goals not the means of human action (Gergely, Nádasdy, Csibra, & Bíró, 1995; Sommerville & Woodward, 2005). By 18 months, toddlers readily imitate the goal of a human's novel, but incomplete action (Meltzoff, 1995). Crucially, an understanding that behaviors are in the service of goals marks this early understanding; toddlers do not slavishly imitate the behavior of a model, but will vary their behavior to achieve the apparent goal of the model.

One study with deaf 4–7-year-olds demonstrated equivalent performance to TH children on a gesture imitation task, with both groups making imitation errors that violated the way in which the hand moved, but remaining faithful to the ultimate goal of the movement trajectory (Want & Gattis, 2005). This pattern of errors indicated that, by 4 years of age, both hearing and deaf children had represented the goal of the action. However, an understanding of intentionality begins in infancy, and we know little of whether deaf infants struggle with an early understanding of intentionality, but overcome this delay by age four.

Several studies have shown that deaf children of hearing parents also show delays in reasoning about desires and intentions. The first showed that deaf children of hearing parents struggle to interpret the intention- and desire-based meanings of eye gaze, failing to correctly infer the desire and intentions represented in a schematic drawing of a face, with only 9-year-olds exhibiting performance similar to that of TH 4-year-olds (Scott, Russell, Gray, Hosie & Hunter, 1999). This delay is quite striking given that pre-linguistic 10-month-olds with normal hearing reliability reliably follow the eye gaze of an experimenter (Brooks & Meltzoff, 2005) and TH 2-year-olds infer desire from eye gaze at above chance levels (Lee, Eskritt, Symons, & Muir, 1998). More low-verbal behavioral tasks conducted with deaf toddlers are needed to fully understand the degree to which language experience affects an understanding of intentionality and desire.

Three other studies that have addressed deaf children's desire reasoning have been in the context of predicting and explaining emotions. Oral deaf children between the ages of 5 and 10 who failed FB tasks also struggled to predict a character's emotion based on their desires, although they were less impaired on the desire-based emotion items than on the FB-based ones (Pyers & de Villiers, 2003). However, deaf 6-year-olds enrolled in a school where emotion reasoning was emphasized in the curriculum were actually more likely to explain emotions in terms of underlying desires than their TH peers (Rieffe & Terwogt, 2000).

In the population of Nicaraguan signers, understanding of desire-based emotions and FB-based emotions were clearly dissociated: failers of a low-verbal FB task readily predicted emotions based on a character's desire in a minimally verbal task (Pyers & de Villiers, 2003; Pyers, 2005). Thus, any limitation in reasoning about desires seems to be overcome before success on FB tasks, and deaf children's understanding of desires as a source of behaviors and emotions appears more robust than their understanding of false beliefs. This is in keeping with the proposal by Wellman (1990) that a desire-based ToM emerges before a belief-based one.

Sources-of-knowledge

Understanding FB is contingent upon understanding how other people's perceptual experiences influence their knowledge states—if the boy does not *see* the chocolate moved from the cupboard to the refrigerator, he does not *know* where it is. Thus, understanding the relationship between sensory perception and knowledge should develop before an understanding of FB (Wellman & Liu, 2004). By age three TH children readily understand Level 1 visual perspective taking—that if someone cannot see the object you can see, that person does not know the identity of the object (Flavell, Everett, Croft, & Flavell, 1981). Yet this understanding does not readily generalize to all senses simultaneously. Children who understand the relationship between seeing and knowing do not seem to exhibit the same understanding of the relationship between feeling and knowing, and they struggle to understand the modality specific nature of knowledge, e.g. that you cannot identify the color of an object by only touching it (O'Neill, Astington, & Flavell, 1992).

Wellman & Liu (2004) included in their ToM scale a “knowledge access” task that taps children's understanding of the relationship between seeing and knowing. Deaf children of hearing parents exhibit delays relative to TH peers on this task, but just like the TH children, they master this concept before they pass traditional FB tasks (Peterson, Wellman, & Liu, 2005; Wellman, Fang, & Peterson, 2011). The “knowledge access” task is a bit more difficult than the traditional seeing-knowing tasks in that it involves complex language, requires the child to inhibit their own knowledge of the object's identity (e.g. Birch & Bloom, 2003), and asks the child to predict another's state of ignorance rather than to report who would be a knowledgeable informant (e.g. O'Neill & Gopnik, 1991; Robinson, Haigh, & Pendle, 2008). Nevertheless, language-delayed deaf children's limitation in understanding seeing and knowing was also observed using a minimally verbal task adapted from Povinelli & de Blois (1992). Deaf children of hearing parents exhibited delays on this task, and their performance correlated with their performance on a low-verbal FB measure and with their language skills (Schick et al., 2007).

Oral deaf children of hearing parents are also delayed relative to TH children on a task adapted from Pratt & Bryant (1990) that taps the seeing-knowing and hearing-knowing relationships. While hearing children succeeded on this task by 4 years of age, deaf children of hearing parents did not pass the task until 5.5 years of age (Schmidt & Pyers, 2011). Most importantly, deaf children were equally delayed on both sensory modalities, even though they have had more limited experience with auditory information relative to visual information. Thus, on a variety of low and high verbal tasks assessing their understanding of the relationship between sensory perception and knowledge, DoH children acquire this understanding well after the age at which typically developing children master it.

Early implicit false belief understanding in indirect tasks

A rapidly growing body of research suggests that in addition to an early understanding of people's goals and intentions, by age 2 or so toddlers have an implicit understanding of others' states

of knowledge and beliefs, even when those beliefs are not in keeping with reality (Baillargeon, Scott, & He, 2010; Low & Perner, 2012). It is not clear how elaborated those concepts are at this age, but they appear to be sufficiently robust to drive toddlers' expectations, selective attention, and anticipatory looking. Some researchers have argued that the difference between this early evidence for understanding of FB and the later emergence of explicit reasoning about FB around age 4 lies in the methods used to assess the children's knowledge. Traditional FB tasks and their variants directly ask about a character's belief or behavior; the research on infants and toddlers uses indirect tests that infer the children's understanding from their eye gaze, anticipatory looking, or spontaneous helping behavior (Baillargeon et al., 2010; Buttelman, Carpenter, & Tomasello, 2009).

Theorists differ on how they regard the difference between children's apparent knowledge on indirect vs. direct tests of FB understanding. Clements & Perner (1994) suggested that anticipatory eye gaze reflected an "implicit" understanding of false belief, and an understanding shown on indirect, but not direct tests of a concept is considered a hallmark of implicit or unconscious knowledge in the literature on consciousness (Low & Perner, 2012). Apperly & Butterfill (2009) suggest there may be two conceptual systems in ToM development: an early-emerging unconscious system that tracks "belief-like" states and is sensitive to another person's engagement with or access to an object or event, and a more abstract, explicit system of concepts with a more articulated representation of the propositional content of beliefs and states of knowledge. Others (e.g. Perner, 2010; Ruffman, Taumoepeau, & Perkins, 2012) have distinguished between the early learning of behavioral regularities or rules that may be more situation-specific, and support expectations and spontaneous behaviors without reference to mental-state representations.

These positions continue to maintain that there is maturation or learning of new conceptual representations in the early preschool years that is reflected in the children's performance on traditional direct tests of FB understanding (Perner, 2010; San Juan & Astington, 2012; Wellman, Cross, & Watson, 2001). In contrast, other researchers argue that the infant and toddler research shows that the core concepts of ToM, including FB, are available to children during the second year of life and may be a part of an innate or early-maturing ToM module (Baillargeon et al., 2010; Leslie & Polizzi, 1998; Roth & Leslie, 1998). However, toddlers cannot make use of that knowledge in the direct, explicit FB reasoning tasks that are used to assess FB because of limited EF resources such as working memory and inhibitory control, skills that are required by those tasks.

The research on EF in deaf children described earlier (de Villiers & de Villiers, 2012; Meristo & Hjelmquist, 2009; Woolfe et al., 2002) indicates that while inhibitory control and set shifting may be prerequisites for success on explicit FB reasoning tasks, those aspects of EF do not seem to be the proximal causes of different levels of performance in those tasks. Language and exposure to fluent communication appear to be the more important proximal predictors. However, we know little to nothing about the early development of EF in deaf children (Marschark, 1993; Marschark & Spencer, 2011).

The research on different aspects of ToM in deaf children that we summarized above all used direct, explicit measures of the children's understanding and reasoning, even when verbal demands of the tasks were minimized. But would language delayed deaf children with hearing parents show early-emerging implicit ToM understanding in the indirect, spontaneous procedures developed for studying TH infants and toddlers? Gale et al. (2009) and de Villiers & de Villiers (2012) demonstrated that on a low-verbal sticker-in-the-hand hide-and-seek game that involved spontaneous deception, language delayed signing and oral deaf children were not delayed in their deceptive behaviors relative to native signing deaf children or matched TH controls, but the youngest children in their studies were 4 years of age.

In a small-scale study of 10 deaf DoH toddlers (17–28 months) with hearing parents Meristo et al. (2012) compared their anticipatory looking in a non-verbal true belief and FB scenario with the same behavior in age-matched TH toddlers matched for age. The parents of the deaf children used spoken Swedish supported by some signs from Swedish Sign Language, but none of the toddlers had mastered many signs. Half of the deaf children had cochlear implants and the other five used hearing aids.

All of the children watched videotaped scenarios in which the familiar cartoon mouse Jerry ran down a tunnel in the shape of a Y and hid in one of two boxes at the end of each arm. He was then followed down the tunnel by the cat Tom. On the test trials Jerry changed his hiding place from one box to the other before Tom entered the tunnel. On the true belief trial Tom was present in the video and saw Jerry move from one box to the other; on the FB trial Tom was not present when the change of location took place. Half of each group of children saw a true belief trial followed by a FB trial; the other half saw the trials in the opposite order. An automated eye tracker measured how long the toddlers watched the exits from each arm of the Y-tunnel or the boxes they led to once Tom entered the tunnel. The TH toddlers spent significantly more time looking at the location that Jerry had moved to in the true belief condition, but significantly more time looking at the empty location (where Tom last saw him) in the FB condition, indicating that their spontaneous attention was sensitive to Tom's state of knowledge. The deaf children also looked more at the correct exit and box in the true belief condition, but in the FB condition they all still looked at the box containing Jerry and none of them looked at the empty box or exit that corresponded to Tom's FB about where he had last seen the mouse.

These deaf toddlers did not show the sensitivity to the character's belief state in spontaneous, anticipatory attention that was demonstrated by the hearing children. Meristo and colleagues suggest two possible interrelated reasons for this result: first, that impoverished communication between the hearing parents and their deaf children impairs the children's acquisition of mental-state concepts or of joint attentional processes that build on the children's gestures and pointing; and second, that early language sharing in hearing caregiver-child dyads may enhance executive functioning abilities that are necessary in this anticipatory looking task where the child has to inhibit the lure of reality and look away from the box containing the mouse. The impoverished communicative interaction between hearing caregivers and their deaf infants and toddlers might delay that EF development. As we point out above, EF does not seem to be delayed at age 4 in language-impaired deaf children with normal range non-verbal IQs and memory development when it is tested in low verbal tasks; but we do not know whether the earliest emerging components of EF are impaired prior to age 3 in these children.

Considering the role of joint attention

DoH children with delayed language acquisition seem to experience broad delays in understanding others' mental states in addition to their difficulties with explicit reasoning about FB. Both joint attention and language may play a causal role in DoH children's impaired ToM, and they both are likely to interact, as language and gestural communication seems to play a crucial role in the development of joint attention (Slaughter, Peterson, & Carpenter, 2009).

Joint attention, the ability to manage attention to both a communication partner and another thing or event, has been posited as a precursor to a full-blown ToM (Moore & Corkum, 1994). Children with autism show delays in joint attention with greater delays in initiating rather than responding to joint attention (Mundy, 2003). DoH children also show some limitations with respect to joint attention, and these limitations may have long-ranging consequences for both their language acquisition and social-cognitive development.

Overwhelmingly, the findings about joint attention in deaf children come out of observational studies of parent-child interaction, and these observational studies show that DoH children have much less experience with joint attention and the quality of that joint attention is significantly poorer than what is observed in typically hearing children. Hearing parents spend less time engaging in coordinated joint attention with their deaf 12- and 18-month-olds than deaf parents with deaf children and hearing parents with hearing TH children (Meadow-Orlans & Spencer, 1996). This limitation in joint attention seems to persist into the second year of life, with deaf 24-month-olds with hearing parents experiencing fewer sustained bouts of joint attention than typically hearing dyads (Gale & Schick, 2009). Another study with slightly different findings showed that DoH children spent more time than hearing children in coordinated joint attention, but the quality of this time was strikingly different. For deaf children, almost none of this time was in symbol-infused joint attention where children were exposed to new words (Prezbindowski, Adamson, & Lederberg, 1998).

That the quantity and quality of joint attention for DoH children is significantly lower than for typically developing children may be the origin of an array of ToM delays. First, joint attention seems to be the hallmark of uniquely human social cognition (Tomasello, Carpenter, Call, Behne, & Moll, 2005). When sharing attention with another person, infants are afforded an opportunity to learn that intentions and goals can be shared (Tomasello, 1995). While the amount of joint attention experience required to support an understanding of shared intentions and goals is unclear, one study comparing hearing infants to deaf infants with and without cochlear implants, showed that more time in joint attention positively correlated with higher maternal perception of social competence (Tasker, Nowakowski, Matilda, & Schmidt, 2010), a finding that is further bolstered by longitudinal studies with typically developing children that have shown that infants' engagement in joint attention at 20 months positively correlates with their performance on a battery of theory of mind measures at 44 months of age (Charman, Baron-Cohen, Swettenham, Baird, Cox, & Drew, 2000). Beyond the amount of joint attention engagement, the quality of joint attention also impacts later ToM development. In a longitudinal study, maternal sensitivity to their infant's internal states at 10 months, a measure that included commenting and elaborating on joint attention, was highly correlated with children's FB scores at 54 months (Ereky-Stevens, 2008).

While experience with joint attention allows for the child to experience a "meeting of the minds," joint attention is where early language learning is situated, and children's engagement in joint attention is predictive of their later language ability (Carpenter, Nagell, & Tomasello, 1998; Mundy & Gomes, 1996; Tomasello & Farrar, 1986). Specifically the degree to which infants respond to bids for joint attention positively correlates with their receptive and productive vocabulary scores at 30 months (Morales, Mundy, Delgado, Yale, Messinger, Neal, & Schwartz, 2000). The relationship between joint attention and language development is present earlier in development: reliably following gaze at 10 months of age predicts vocabulary scores at 18 months (Brooks & Meltzoff, 2005). The early impact that joint attention has on vocabulary development likely also has consequences for more complex language development, such that the language delays faced by deaf children because of their hearing-impairment may be compounded by their limited joint attention experience.

The development of ToM and of language is dependent upon extensive experience with high-quality, symbol-infused joint attention. As such, the ToM delays observed in most deaf children of hearing families may originate in the hearing parents' struggle to engage in joint attention with their deaf children; none of the observational studies of deaf children observed children's failure to point or to follow eye-gaze as has been observed for children with autism. Instead, deaf children exhibit the appropriate joint attention behaviors, but they are given limited opportunity

to engage in them. This limitation has a two-fold impact on deaf children's ToM development. First, what is seen as one of the key precursors to a mature ToM is impoverished, likely impacting later developments in ToM reasoning. Secondly, the limited experience with joint attention can delay language development and language is highly predictive of successful FB reasoning during the preschool years (see Milligan, Astington, & Dack, 2007 for a review). Thus, a key locus of ToM intervention for deaf children with hearing families may be in teaching hearing parents how to initiate and sustain high-quality joint attention with their deaf children who have the capacity, but limited opportunity, to do so.

Conclusions

The case of DoH children sheds light on the way in which environment, specifically language experience, shapes the development of a mature ToM. The research on ToM in deaf children that we have summarized in this chapter leads us to several conclusions:

1. Language acquisition and comprehensible communication from infancy are essential for the development of ToM on a normal timetable. DoH children with language delays and impoverished communicative interactions are significantly delayed in several aspects of their ToM development.
2. Initial evidence suggests that this delay in ToM includes not only explicit, propositional reasoning about FB and states of knowledge in preschoolers, but also implicit, spontaneous expectations and anticipatory behaviors in infants and toddlers. Much more research is needed on these early stages of ToM understanding in both language-delayed deaf toddlers and native signing deaf children.
3. Any impairment in communication between hearing caregivers and their deaf children may lead to failures in initiating and sustaining high-quality joint attention with their deaf infants may contribute to both further language acquisition delays and to impaired understanding of the mental states of others.
4. The degree of impairment seen in explicit ToM tasks is most strongly predicted by the acquisition of complex language by the deaf children and by the degree of language delay that they experience. Early interventions with a comprehensible natural sign language or effective amplification systems enhance spoken language acquisition (Rommel & Peters, 2009) can considerably mitigate these delays in social cognition.
5. Development of a fully articulated ToM seems to not be constrained by a critical period, and can continue well into later childhood or even adulthood (Peterson, 2009; Pyers & Senghas, 2009; Wellman et al., 2011). Thus, the research on the development of ToM in deaf children has not only led to a better understanding of the importance of language experience in that development, it has argued against any strong critical period for ToM development and provided some optimism for the effectiveness of interventions to facilitate deaf children's social cognitive development.

References

- Apperly, I., & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review* 20: 521–36.
- Baillargeon, R., Scott, R., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Science* 14: 110–18.

- Bartsch, K., & Wellman, H. (1989). Young children's attribution of action to beliefs and desires. *Child Development* 60(4): 946–64.
- Birch, S. J., & Bloom, P. (2003). Children are cursed: An asymmetric bias in mental-state attribution. *Psychological Science* 14(3): 283–6.
- Brooks, R., & Meltzoff, A. N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology* 38(6): 958–66.
- Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science* 8(6): 535–43.
- Buttelman, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition* 112: 337–42.
- Carlson, S., & Moses, L. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development* 72: 1032–53.
- Carlson, S., Moses, L., & 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.
- 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).
- Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Cox, A., & Drew, A. (2000). Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development* 15(4): 481–98.
- Cicchetti, D., Rogosch, F. A., Maughan, A., Toth, S. L., & Bruce, J. (2003). False belief understanding in maltreated children. *Development and Psychopathology* 15(4): 1067–91.
- Clements, W., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development* 9: 377–95.
- Cokely, D. (1983). When is a pidgin not a pidgin? An alternate analysis of the ASL-English contact situation. *Sign Language Studies* 38 1–24.
- Corina, D., & Singleton, J. (2009). Developmental social cognitive neuroscience: Insights from deafness. *Child Development* 80(4): 952–67.
- Courtin, C. (2000). The impact of sign language on the cognitive development of deaf children: The case of theories of mind. *Journal of Deaf Studies and Deaf Education* 5(3): 266–76.
- Davis, H. & Pratt, C. (1995). The development of children's theory of mind: The working memory explanation. *Australian Journal of Psychology* 47: 25–31.
- de Villiers, J. G. & de Villiers, P. A. (2009). Complements enable representation of the contents of false beliefs: The evolution of a theory. In S. Foster-Cohen (Ed.), *Language acquisition* (pp. 169–95). Basingstoke: Palgrave-McMillan 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: 1037–60.
- de Villiers, P. A. (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. 266–97). New York: Oxford University Press.
- de Villiers, P. A., & de Villiers, J. G. (2012). Deception dissociates from false belief reasoning in deaf children: Implications for the implicit vs. explicit theory of mind distinction. *British Journal of Developmental Psychology* 30(1): 188–209.
- de Villiers, P. A. & Pyers, J. (2001). Complementation and false-belief representation. In M. Almgren, A. Barrena, M. J. Ezeizabarrena, I. Idiazabal, & B. MacWhinney (Eds.), *Research on Child Language Acquisition: Proceedings of the 8th Conference of the International Association for the Study of Child Language*. (pp. 984–1005). Somerville: Cascadia Press.
- Edmondson, P. (2006). Deaf children's understanding of other people's thought processes. *Educational Psychology in Practice* 22: 159–69.

- Engen, E. & Engen, T. (1983). *The Rhode Island Test of Language Structure* (RITLS). Austin: Pro-Ed.
- Ereky-Stevens, K. (2008). Associations between mothers' sensitivity to their infants' internal states and children's later understanding of mind and emotion. *Infant and Child Development* 17(5): 527–43.
- Figueras-Costa, B., & Harris, P. (2001). Theory of mind development in deaf children: A nonverbal test of false-belief understanding. *Journal of Deaf Studies and Deaf Education* 6(2): 92–102.
- Flavell, J. H. (1992). Perspectives on perspective taking. In H. Beilin & P. B. Pufall (Eds), *Piaget's Theory: Prospects and Possibilities* (pp. 107–39). London: Psychology Press.
- Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology* 17(1): 99–103.
- Frye, D., Zelazo, P., & Palfrai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development* 10: 483–527.
- Gale, E., de Villiers, P., de Villiers, J., & Pyers, J. (1996). Language and theory of mind in oral deaf children. In A. Stringfellow, D. Cahana-Amitay, E. Hughes, & A. Zukowski (Eds), *Proceedings of the 20th Annual Boston University Conference on Language Development*, Vol. 1. (pp. 213–24). Somerville: Cascadilla Press.
- Gale, E., de Villiers, P., Schick, B., Hoffmeister, R., & Pyers, J. (2009). *Deception in Oral and Signing Deaf Children: Not Delayed nor Dependent on Complex Language*. Poster presented at the biennial meeting of the Society for Research in Child Development, Denver, CO.
- Gale, E., & Schick, B. (2009). Symbol-infused joint attention and language use in mothers with deaf and hearing toddlers. *American Annals of the Deaf* 153(5): 484–503.
- Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition* 56(2): 165–93.
- Jackendoff, R. (1996). How language helps us think. *Pragmatics and Cognition* 4: 1–34.
- Keenan, T. (2000). Mind, memory, and metacognition: The role of memory span in children's developing understanding of the mind. In J. W. Astington (Ed.), *Minds in the Making: Essays in Honor of David R. Olson* (pp. 233–49). Oxford: Blackwell Publishers.
- Lee, K., Eskritt, M., Symons, L. A., & Muir, D. (1998). Children's use of triadic eye gaze information for “mind reading.” *Developmental Psychology* 34(3): 525–39.
- Leslie, A., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science* 1: 247–54.
- Lieberman, A., Hatrak, M., & Mayberry, R. I. (2011). The development of eye gaze control for linguistic input in deaf children. In N. Danis, K. Mesh, & H. Sung, (Eds), *Proceedings of the 35th Annual Boston University Conference on Language Development* (pp. 391–404). Somerville: Cascadilla Press.
- Low, J., & Perner, J. (2012). Implicit vs. explicit theory of mind: State of the art. *British Journal of Developmental Psychology* 30: 1–13.
- Lucas, C., & Valli, C. (1989). Language contact in the American deaf community. In C. Lucas (Ed.), *The Sociolinguistics of the Deaf Community* (pp 11–40). San Diego: Academic Press.
- Marschark, M. (1993). *Psychological Development of Deaf Children*. New York: Oxford University Press.
- Marschark, M. & Spencer, P. (Eds) (2011). *The Oxford Handbook of Deaf Studies, Language and Education*, Vol. 1, 2nd edn. New York: Oxford University Press.
- Mayberry, R. I., & Lock, E. (2003). Age constraints on first vs. second language acquisition: Evidence for linguistic plasticity and epigenesis. *Brain and Language* 87(3): 369–84.
- Meadow-Orlans, K. P., & Spencer, P. E. (1996). Maternal sensitivity and the visual attentiveness of children who are deaf. *Early Development & Parenting* 5(4): 213–23.
- Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology* 31(5): 838–50.
- Meristo, M., Falkman, K. W., Hjelmquist, E., Tedoldi, M., Surian, L., & Siegal, M. (2007). Language access and theory of mind reasoning: Evidence from deaf children in bilingual and oralist environments. *Developmental Psychology* 43(5): 1156–69.

- Meristo, M., & Hjelmquist, E. (2009). Executive functions and theory-of-mind among deaf children: Different routes to understanding other minds? *Journal of Cognition and Development* 10(1–2): 67–91.
- Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L., & Siegal, M. (2012). Belief attribution in deaf and hearing infants. *Developmental Science* 15: 1–9.
- Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development* 78, 622–46.
- Mitchell, R. E. (2006). How many deaf people are there in the United States? Estimates from the survey of income and program participation. *Journal of Deaf Studies and Deaf Education* 11(1): 112–19.
- Moeller, M. P., & Schick, B. (2006). Relations between maternal input and theory of mind understanding in deaf children. *Child Development* 77(3): 751–66.
- Moore, C., & Corkum, V. (1994). Social understanding at the end of the first year of life. *Developmental Review* 14(4): 349–72.
- Morgan, G., & Kegl, J. (2006). Nicaraguan Sign Language and theory of mind: The issue of critical periods and abilities. *Journal of Child Psychology and Psychiatry* 47(8): 811–19.
- Moses, L. (2001). Executive accounts of theory-of-mind development. *Child Development* 72: 688–90.
- Mundy, P. (2003). The neural basis of social impairments in autism: the role of the dorsal medial-frontal cortex and anterior cingulate system. *Journal of Child Psychology and Psychiatry* 44(6): 793–809.
- Mundy, P., & Gomes, A. (1998). Individual differences in joint attention skill development in the second year. *Infant Behavior and Development* 21: 468–82.
- O'Neill, D. K., Astington, J. W., & Flavell, J. H. (1992). Young children's understanding of the role that sensory experiences play in knowledge acquisition. *Child Development* 63(2): 474–90.
- O'Neill, D. K., & Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology* 27(3): 390–7.
- Perner, J. (2010). Who took the cog out of cognitive science: Mentalism in an era of anti-cognitivism. In P. A. Frensch & R. Schwarzer (Eds), *Cognition and Neuropsychology: Proceedings of the 29th International Congress of Psychology*, Vol. 1. London: Psychology Press.
- Peterson, C. C. (2009). Development of social-cognitive and communication skills in children born deaf. *Scandinavian Journal of Psychology* 50: 475–83.
- Peterson, C. C., & Siegal, M. (1995). Deafness, conversation and theory of mind. *Journal of Child Psychology and Psychiatry* 36(3): 459–74.
- Peterson, C. C., Wellman, H., & Liu, D. (2005). Steps in theory of mind development among children with autism, deafness or typical development. *Child Development* 76: 502–17.
- Povinelli, D., & de Bois, S. (1992). Young children's (*Homo sapiens*) understanding of knowledge formation in themselves and others. *Journal of Comparative Psychology* 106(3): 228–38.
- Povinelli, D., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee mind. *Mind and Language* 19: 1–28.
- Pratt, C., & Bryant, P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development* 61(4): 973–82.
- Prezbindowski, A. K., Adamson, L. B., & Lederberg, A. R. (1998). Joint attention in deaf and hearing 22 month-old children and their hearing mothers. *Journal of Applied Developmental Psychology* 19(3): 377–87.
- Pyers, J. (2005). The relationship between language and false-belief understanding: Evidence from learners of an emerging sign language in Nicaragua. *Dissertation Abstracts International*, 66.
- Pyers, J., & de Villiers, P. (2003). Theory of mind and understanding emotions: What is the role of complex language? Poster session presented at the biennial meeting of the Society for Research in Child Development, Tampa, FL.
- Pyers, J., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from Nicaraguan sign language. *Psychological Science* 20: 805–12.

- Rommel, E., & Peters, K. (2009). Theory of mind and language in children with cochlear implants. *Journal of Deaf Studies and Deaf Education* 14, 218–36.
- Rieffe, C., & Terwogt, M. (2000). Deaf children's understanding of emotions: Desires take precedence. *Journal of Child Psychology and Psychiatry* 41(5): 601–8.
- Robinson, E. J., Haigh, S. N., & Pendle, J. C. (2008). Children's working understanding of the knowledge gained from seeing and feeling. *Developmental Science* 11(2): 299–305.
- Roth, D., & Leslie, A. (1998). Solving belief problems: Toward a task analysis. *Cognition* 66, 1–31.
- Ruffman, T., Taumoepeau, M., & Perkins, C. (2012). Statistical learning as a basis for social understanding in children. *British Journal of Developmental Psychology* 30: 59–74.
- Russell, P., Hosie, J., Gray, C., Scott, C., Hunter, N., Banks, J. & Macaulay, M. (1998). Development of theory of mind in deaf children. *Journal of Child Psychology and Psychiatry* 39: 903–10.
- San Juan, V., & Astington, J. (2012). Bridging the gap between implicit and explicit understanding: How language development promotes the processing and representation of false belief. *British Journal of Developmental Psychology* 30: 105–22.
- Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and theory of mind: A study of deaf children. *Child Development* 78(2): 376–96.
- Schick, B., Williams, K., & Kupermintz, H. (2006). Look who's being left behind: Educational interpreters and access to education for deaf and hard-of-hearing students. *Journal of Deaf Studies and Deaf Education* 11(1): 3–20.
- Schick, B. (2004). How might learning through an educational interpreter influence cognitive development? In E. A. Winston (Ed.), *Educational Interpreting: How It Can Succeed* (pp. 73–87). Washington, DC: Gallaudet University Press.
- Schmidt, E., & Pyers, J. (2011). Children's understanding of the link between sensory perception and knowledge. In L. Carlson, C. Hoelscher, & T. Shipley (Eds), *The Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 3016–21). Austin: Cognitive Science Society.
- Scott, C., Russell, P. A., Gray, C. D., Hosie, J. A., & Hunter, N. (1999). The interpretation of line of regard by prelingually deaf children. *Social Development* 8(3): 412–26.
- Shatz, M., Diesendruck, G., Martinez-Beck, I., & Akar, D. (2003). The influence of language and socioeconomic status on children's understanding of false belief. *Developmental Psychology* 39(4): 717–29.
- Slaughter, V., Peterson, C. C., & Carpenter, M. (2009). Maternal mental state talk and infants' early gestural communication. *Journal of Child Language* 36: 1053–74.
- Sommerville, J. A., & Woodward, A. L. (2005). Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition* 95(1): 1–30.
- Tasker, S. L., Nowakowski, M. E., & Schmidt, L. A. (2010). Joint attention and social competence in deaf children with cochlear implants. *Journal of Developmental and Physical Disabilities* 22(5): 509–32.
- Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. Dunham (Eds), *Joint Attention: Its Origins and Role in Development*. (pp. 85–101). Hillsdale: Erlbaum.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* 28(5): 675–735.
- Tomasello, M., & Farrar, M. Y. (1986). Joint attention and early language. *Child Development* 57: 1454–63.
- Vaccari, C., & Marschark, M. (1997). Communication between parents and deaf children: Implications for social-emotional development. *Journal of Child Psychology and Psychiatry* 38: 793–801.
- Want, S. C., & Gattis, M. (2005). Are "late-signing" deaf children "mindblind"? Understanding goal directedness in imitation. *Cognitive Development* 20(2): 159–72.
- Wellman, H. (1990). *The Child's Theory of Mind*. Cambridge: MIT Press.

- Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72: 655–84.
- Wellman, H., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory-of-mind scale: Longitudinal perspectives. *Child Development* 82(3): 780–92.
- Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development* 75(2): 523–41.
- Woolfe, T., Want, S. C., & Siegal, M. (2002). Signposts to development: Theory of mind in deaf children. *Child Development* 73(3): 768–78.

Social cognition in individuals with psychopathic tendencies

James Blair and Stuart F. White

The goal of this chapter is to consider social cognition in individuals with psychopathic tendencies. As such, we will first consider the nature of psychopathy. Following this, we will briefly outline our position on psychopathy: the integrated emotion systems (IES) model. This model will be used as the backdrop for considering social cognition in individuals with psychopathic tendencies. Specifically, the model advocates a multi-system approach to social cognition with computationally distinct architectures proposed for cognitive empathy (theory of mind), motor empathy and different forms of emotional empathy (responses to sad/fear, responses to pain, responses to disgust, responses to anger). The ability of individuals with psychopathic tendencies to show these forms of social cognition will be considered in turn. The chapter will then conclude with a consideration of a developmental consequence of the dysfunctional empathy/emotional systems seen in psychopathy, i.e. a profound disruption in moral judgment.

Psychopathy

The disorder of **psychopathy** characterizes an individual who shows pronounced emotional deficits and is at increased risk for displaying antisocial behavior (Frick, 1995; Hare, 2003). The emotional deficits present clinically as reduced guilt and “empathy” (callous and unemotional (CU) traits). The disorder is developmental. It has been shown that CU traits in particular and the disorder more generally are relatively stable from childhood into adulthood (Lynam, Caspi, Moffitt, Loeber, & Stouthamer-Loeber, 2007; Munoz & Frick, 2007). In addition, the functional impairments seen in adults with psychopathy (e.g. in responding to emotional expressions, aversive conditioning, passive avoidance learning, reversal learning, extinction) are also seen in adolescents with psychopathic tendencies.

Assessment scales for psychopathy include the Antisocial Process Screening Device (Frick & Hare, 2001) and Psychopathy Checklist—Youth Version (Forth, Kosson, & Hare, 2007) for adolescents and the Psychopathy Checklist-Revised (Hare, 2003) for adults. These typically identify three dimensions of behavior (Cooke, Michie, & Hart, 2006; Frick, Bodin, & Barry, 2000; Neumann, Kosson, Forth, & Hare, 2006) though the exact number is debated (Cooke et al., 2006). These three dimensions include:

1. An emotional factor focusing on CU traits
2. An arrogant and deceitful interpersonal style involving narcissism.
3. Impulsive and irresponsible behavior (Cooke et al., 2006; Frick et al., 2000; Neumann et al., 2006).

Three (or more) factor solutions of psychopathy assessment measures do not imply that the disorder involves a triad of impairments. The strong assumption underlying this chapter is that there is a single underlying impairment that gives rise to the presence of CU traits and that this increases the risk for antisocial behavior (and presumably narcissism, although this latter relationship remains underspecified). Neurocognitive impairments may be identified that are particularly associated with each of the three factors. This does not mean, however, that psychopathy involves all of these impairments. Indeed, impairments in “cold” (non-affect driven) executive functions—those relating to working memory and “inhibitory control” are associated with an increased risk for impulsive behavior (Moffitt, 1993; Seguin, Boulerice, Harden, Tremblay, & Pihl, 1999). However, individuals with psychopathy, as youths or adults, show no significant impairment in these forms of executive function (Blair, Newman, Mitchell, Richell, Leonard, Morton, et al., 2006; Hart, Forth, & Hare, 1990; LaPierre, Braun, & Hodgins, 1995). The suggestion is that psychopathy is associated with elevated CU traits and that these traits put the individual at increased risk for antisocial behavior. However, individuals can also be at risk for (more impulsive) antisocial behavior if they show “cold” executive dysfunction. In short, there are many developmental routes to an elevated risk for antisocial behavior (Blair, 2004; Frick & Marsee, 2006).

It should also be noted that the disorder of psychopathy is not equivalent to the DSM-IV diagnoses of conduct disorder or antisocial personality disorder (ASPD) or their ICD-10 counterparts. These psychiatric diagnoses concentrate on the presence of antisocial behaviors, rather than underlying causes, such as the emotion dysfunction seen in psychopathy (Blair, Mitchell, & Blair, 2005). As such, they capture individuals whose difficulties relate to executive dysfunction (Moffitt, 1993), as well as individuals whose difficulties relate to CU traits. As a consequence, individuals meeting the criteria for conduct disorder and antisocial personality disorder are more heterogeneous in their pathophysiology than individuals meeting criteria for psychopathy (Karnik, McMullin, & Steiner, 2006). It should be noted, however, that DSM-5 looks likely to introduce a CU specifier when considering the diagnosis of conduct disorder (CD) and that the diagnosis of ASPD will include components of psychopathy. While the disorder of psychopathy will still not be equivalent to the DSM diagnoses of CD and ASPD, there will be closer relationship of these conceptualizations.

The integrated emotion systems model

The IES model provides a cognitive neuroscience perspective on psychopathy (Blair, 2007). Core theoretical components of this model are:

1. Emotional expressions serve as reinforcers. Actions associated with the reward of another individual's happiness will be represented as “good.” In contrast, actions associated with the punishment of another individual's sadness or fear will be represented as “bad.”
2. There are relatively independent emotion learning systems that process the reinforcement provided by specific emotional expressions; the amygdala is particularly associated with the processing of fearful and sad expressions (i.e. activation of the amygdala by such expressions initiates stimulus-reinforcement learning such that representations of objects/actions associated with these expressions acquire valence), the insula is associated with disgusted expressions, and the inferior frontal cortex with angry expressions. It is as a direct developmental consequence of these emotional learning systems that humans have developed multiple moralities: care-based (actions associated with harming/helping others), disgust-based (actions associated with the disgust of others) and conventional (actions associated with hierarchical violations leading to interpersonal anger).

3. Emotional learning systems feed reinforcement expectancy information to ventromedial prefrontal cortex (vmPFC). vmPFC represents this information allowing successful decision making, including moral decision making. This is particularly important in situations when it is necessary to choose between response options.
4. These neural systems are critical when deciding upon behavioral choices. They are also critical because they allow moral transgressions and prosocial actions to gain emotive force.

According to the IES model, the roles of the amygdala in stimulus-reinforcement learning, the caudate in prediction error signaling and vmPFC in the representation of reinforcement expectancy information are all disrupted in psychopathy (Blair, 2007; Finger, Marsh, Blair, Reid, Sims, Ng, et al., 2011). In contrast, the roles of insula and inferior frontal cortex in response to disgusted and angry expressions are not. Importantly, this is a cognitive neuroscience model—the emphasis is on functional roles within neural systems. There is no assumption that because the role of the amygdala in stimulus-reinforcement learning is disrupted, all functions of the amygdala are disrupted. Similarly, there is no assumption that because the insula's response to disgust information is intact, all functions of the insula are intact.

Social cognition

Social cognition involves the processing of information relating to conspecifics in the brain. Aspects of social cognition that will be concentrated on in this chapter include the different processes subsumed under the term empathy and some of the processes consequent on, or related to, them; i.e. (social) decision making and moral reasoning.

Empathy can be defined as “an affective response more appropriate to someone else's situation than to one's own” (Hoffman, 1987, p. 48). Empathy has been considered a unitary process. Thus, Preston and de Waal (2002) have argued that “empathy [is] a super-ordinate category that includes all sub-classes of phenomena that share the same mechanism. This includes emotional contagion, sympathy, cognitive empathy, helping behavior, etc.” (Preston & de Waal, 2002, p. 4). However, strong arguments have been put forward against unitary models of empathy in which all classes of the phenomenon *share the same mechanism* (Blair, 2005; Blair, 2006) and evidence against unitary models is growing (Decety, 2011). According to the alternative view, at least three main divisions of social cognitive process that are generally referred to as empathy can be distinguished: cognitive, motor, and emotional empathy (Blair, 2005; Blair, 2006). These rely on partial overlapping neural architectures (Blair, 2005; Blair, 2006). Cognitive empathy occurs when the individual represents the internal mental state of another individual. As such it reflects the narrower definitions of ToM (Baron-Cohen, Leslie, & Frith, 1985). Motor empathy occurs when the individual mirrors the motor responses of the observed actor. Emotional empathy reflects emotional responses to emotional social cues (expressive displays of affect and pain, but also verbal stimuli, such as the phrase “Adam just lost his house”).

The current chapter will briefly consider the ability of individuals with psychopathic traits to perform these three distinct empathic processes as well as a major developmental consequence of them—moral reasoning.

Cognitive empathy (theory of mind) and psychopathic traits

Theory of mind (ToM) refers to the ability to represent the mental states of others, i.e. their thoughts, desires, beliefs, intentions, and knowledge (Frith, 1989). ToM allows the attribution of

mental states to self and others in order to explain and predict behavior. Neural regions considered critical for mediating theory of mind include medial frontal cortex, temporal parietal junction, posterior cingulate cortex, and temporal pole (Amodio & Frith, 2006; Lombardo, Chakrabarti, Bullmore, Wheelwright, Sadek, Suckling, et al., 2010; Saxe & Baron-Cohen, 2006).

There are no indications of ToM impairment in individuals with psychopathy. Six out of seven studies assessing the ability of individuals with psychopathic tendencies on ToM measures have reported no impairment (Blair, Sellars, Strickland, Clark, Williams, Smith, et al., 1996; Dolan & Fullam, 2004; Jones, Happe, Gilbert, Burnett, & Viding, 2010; Richell, Mitchell, Newman, Leonard, Baron-Cohen, & Blair, 2003; Shamay-Tsoory, Harari, Aharon-Peretz, & Levkovitz, 2010; Widom, 1978). Only one study has reported impairment and this used a rating scale that is not a typical measure of ToM (Widom, 1976).

People with autism have been shown to present with impairment in ToM (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Frith, 1989). Interestingly, several studies with individuals with psychopathic tendencies have used measures that demonstrate ToM impairment in individuals with autism (Blair et al., 1996; Dolan & Fullam, 2004; Jones et al., 2010; Richell et al., 2003). Yet none of these studies revealed ToM impairment in individuals with psychopathic tendencies.

These findings are consistent with suggestions that functions mediated by neural regions implicated in ToM (medial frontal cortex, temporal parietal junction, posterior cingulate cortex and temporal pole) are not, with the possible exception of posterior cingulate cortex, considered impaired in psychopathy (Blair, 2007; Finger et al., 2011). Kiehl has proposed that the functions of all of cingulate cortex is disrupted in psychopathy (Kiehl, 2006). We would argue that the ToM data indicates that at least some functions of posterior cingulate cortex remain intact.

Three of the studies did report subtle impairments on tasks involving the representation of the **emotional** states of others (Dolan & Fullam, 2004; Jones et al., 2010; Shamay-Tsoory et al., 2010). In perhaps the most interesting of these, Shamay-Tsoory and colleagues (2010) demonstrated that these emotional deficits closely followed those found in patients with orbital frontal cortex damage. Orbital frontal cortex, and its role in the representation of emotional information, is thought to be dysfunctional in psychopathy (Blair, 2007).

Motor empathy and psychopathic traits

Motor empathy is defined as the tendency to automatically mimic and synchronize facial expressions, vocalizations, postures, and movements with those of another person (Hatfield, Cacioppo, & Rapson, 1994). A neurocognitive account of motor empathy has been developed (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003; Decety & Jackson, 2004). This account relied heavily on the discovery of mirror neurons—neurons that show activity during the execution and also the observation of an action (Rizzolatti, Fogassi, & Gallese, 2001). Human fMRI work has suggested a commonality between regions involved in action execution and those activated when actions are observed (Blakemore & Decety, 2001).

Within the neurocognitive account of motor empathy, the perception of another individual's state activates the observer's corresponding representations, which in turn activate somatic and autonomic responses. At the anatomical level, superior temporal cortex, posterior parietal and inferior frontal cortex have been implicated (Carr et al., 2003). Connections from these regions to the insula are then thought to allow this representation information to generate emotional responses through limbic areas (Carr et al., 2003).

At one stage, this mirror neuron based account was a popular model of all empathy supposedly allowing cognitive, motor and emotional empathy to occur (Carr et al., 2003; Decety & Jackson,

2004; Keysers & Gazzola, 2006). However, this approach is open to considerable criticism (Blair, 2005, 2006) and at least some early adoptees of the view now place considerably less reliance on the construct in their theorizing (Decety, 2011). While mirror neuron based accounts struggle to account for all aspects of empathy, they are extremely useful when considering motor empathy—the automatic mimicking and synchronizing of facial expressions, vocalizations, postures, and movements with those of another person. Indeed, it is not difficult to see how the activity of mirror neurons could underpin motor empathy. The sight of another committing an act would prime those neurons mediating the act in the viewer and thus increase the probability that the viewer would elicit the act in the presence of other environmental stimuli triggering its display.

What about motor empathy in psychopathy, however? Partly because mirror neurons have not been seriously considered as the basis of an account of the disorder, no real work has investigated the issue in this population. Of the regions implicated in the human mirror neuron system (superior temporal cortex, posterior parietal and inferior frontal cortex; Carr et al., 2003), only activity within superior temporal cortex is consistently observed to be dysfunctional in individuals with psychopathic tendencies (Blair, 2010). While there is an account that considers the functioning of superior temporal cortex to be compromised in psychopathy (Kiehl, 2006), and thus should predict that the mirror neuron system is compromised in this population, there are no data to support the idea. Moreover, if mirror neuron system based accounts of autism prove to be useful (Thioux & Keysers, 2010), it is unlikely that they will be useful as the basis of accounts of psychopathic tendencies given the marked differences in the pattern of impairment between these two conditions (Blair, 2008). In short, the existing data suggests that motor empathy, like cognitive empathy/ToM, is intact in individuals with psychopathic tendencies.

Emotional empathy and psychopathic traits

As noted above, emotional empathy reflects emotional responses to emotional social cues (emotional expressions). As such, it is important to consider what function emotional empathy serves. We have argued that facial expressions of emotion have specific communicatory functions, that they impart specific information to the observer (Blair, 2003; see also; Fridlund, 1992). From this point of view, empathy to facial and vocal emotional expressions is effectively the “translation” of the communication by the observer. Moreover, because of the different implications of these communicatory signals, they are translated in several separable neural systems (Blair, 2003). More specifically, they prompt learning; the individual associates the valence of the emotional expression so that the object/action the individual is displaying the emotional reaction to comes to have this valence. The idea is that this forms the basis of (different forms of) moral judgment (Blair, 2007).

Processing the fear and sadness of others

Fearfulness and sadness can be viewed as aversive reinforcers that reduce the probability that a particular behavior will be performed in the future (Blair, 2003). Indeed, observational fear studies in animals have shown that fearful faces can be considered to be aversive unconditioned stimuli that rapidly convey information to others that a novel stimulus is aversive and should be avoided (Mineka & Cook, 1993). Similarly, it has been argued that sad facial expressions also act as aversive unconditioned stimuli discouraging actions that caused the display of sadness in another individual and motivating reparatory behaviors (Blair, 1995).

The amygdala has been implicated in aversive (and appetitive) conditioning including instrumental learning (Cardinal & Everitt, 2004; LeDoux, 2007). If fearful and sad expressions induce aversive conditioning (they can be considered aversive unconditioned stimuli that when associated

with novel stimuli will provide negative valence to these novel stimuli), then it is to be expected that the amygdala will be importantly involved in their processing (Blair, 2003). In line with this, amygdala lesions have been consistently associated with impairment in the recognition of fearful expressions (Adolphs, 2002). While less robustly observed than the fear recognition impairment, impairment in the recognition of sad expressions is not uncommonly found in patients with amygdala lesions (Adolphs & Tranel, 2004; Schmolck & Squire, 2001). Indeed, a review of patient performance across studies reported that approximately half of all patients with amygdala damage present with impairment for the recognition of sad expressions (Fine & Blair, 2000). Strikingly, and highly consistent with the above, recent animal work has shown that amygdala lesions block the acquisition and expression of observational fear (Jeon, Kim, Chetana, Jo, Ruley, Lin, et al., 2010).

Both children and adults with psychopathic tendencies relatively consistently show impairment in the recognition of fearful and sad facial and vocal affect (Blair, Colledge, Murray, & Mitchell, 2001b; Blair, Mitchell, Richell, Kelly, Leonard, Newman, et al., 2002; Dadds, Perry, Hawes, Merz, Riddell, Haines, et al., 2006; Dolan & Fullam, 2006; Stevens, Charman, & Blair, 2001). Indeed, a meta-analytic review of the field reported a robust link between psychopathy/antisocial behavior and specific deficits in the recognition of fearful expressions (Marsh & Blair, 2008). Moreover, both adults with psychopathy and children with psychopathic tendencies show reduced autonomic responses (Blair, 1999; Blair, Jones, Clark, & Smith, 1997) and reduced attention to (Kimonis, Frick, Fazekas, & Loney, 2006) the sad expressions of others. Importantly, it was these findings, and previous work demonstrating that augmentation of the startle reflex (a function in which the amygdala plays a critical role) was reduced in psychopathy, that first lead to the suggestion of amygdala dysfunction in this population (Blair, Morris, Frith, Perrett, & Dolan, 1999; Patrick, 1994).

One interesting feature to note here is that the amygdala does not only allow aversive conditioning, it also plays a role in attention, priming stimulus features that have emotional content (Gallagher & Schoenbaum, 1999). This is particularly interesting with respect to fearful expressions as the amygdala plays a role in directing gaze toward features of the face that are particularly important for the recognition of the fearful expression, notably the eyes. Patients with amygdala damage show reduced gaze toward the eyes of individuals displaying fear and show a reduction in the fear recognition deficit when instructed to attend to the eyes (Adolphs, Gosselin, Buchanan, Tranel, Schyns, & Damasio, 2005). In a series of studies, Dadds and colleagues have shown that youth with psychopathic traits similarly direct gaze less toward the eyes (Dadds, El Masry, Wimalaweera, & Guastella, 2008) and show improvement in their fear recognition following instructions to attend to the eyes (Dadds et al., 2006). There have been suggestions that this impact of the amygdala on gaze direction “explains” the fear recognition deficit—that the impairment in psychopathic tendencies is in attending to the facial features of the fearful expression rather than the emotional response to these features (Dadds et al., 2008). However, recent data shows that this interpretation is incorrect. Using a continuous flash suppression paradigm in which fearful or neutral faces were presented to the non-dominant eye and a scrambled image was presented to the dominant eye so that participants were only conscious of the scrambled image, Lilienfeld and colleagues demonstrated that recognition deficits for fearful expressions were present even under presentation conditions too rapid for group differences in gaze direction to have an impact (Sylvers, Brennan, & Lilienfeld, 2011). As such, the suggestion is that the amygdala’s response to fearful and sad expressions is disrupted in psychopathy and that this in turn leads to reduced gaze direction to emotional salient facial features, presumably exacerbating the reduction relative to healthy individuals, in responding to these expressions.

Functional neuroimaging data has generally supported the neuropsychological data. Certainly, the amygdala is significantly more responsive to fearful than to other emotional expressions (see,

for a meta-analytic review, Murphy, Nimmo-Smith, & Lawrence, 2003). Less work has examined the neural response to sad expressions, but amygdala responses to this expression are also reported (e.g. Blair et al., 1999; Drevets, Lowry, Gautier, Perrett, & Kupfer, 2000). With respect to fMRI work with youth with psychopathic tendencies, reduced amygdala responses have been shown to fearful expressions (Jones, Laurens, Herba, Barker, & Viding, 2009; Marsh, Finger, Mitchell, Reid, Sims, Kosson, et al., 2008) and, in youth with conduct disorder, to sad expressions (Passamonti, Fairchild, Goodyer, Hurford, Hagan, Rowe, et al., 2010). Less work has been conducted with adults, but a non-clinical sample showed reduced amygdala responses to fearful expressions as a function of psychopathy (Gordon, Baird, & End, 2004); however, it should be noted that another study with a clinical sample did not (Pardini & Phillips, 2010). In short, the response to the sadness and fear of others is disrupted in individuals with psychopathic traits.

Processing the pain of others

Responding to another in pain most typically activates the insula and dorsomedial frontal cortex (Jackson, Rainville, & Decety, 2006; Singer & Lamm, 2009), although other regions such as the amygdala, striatum and periaqueductal gray also appear to be implicated (Decety, 2011). The literature on the response to pain has typically viewed the response to the pain of others in terms of the activation of shared representations between the self and other that allow the individual to consciously feel the emotion state (within this literature, the pain) of the other individual (Bastiaansen, Meffert, Hein, Huizinga, Ketelaars, Pijnenborg, et al., 2011; Jackson et al., 2006; Singer & Lamm, 2009). Indeed, the literature suggests that at least insula and dorsomedial frontal cortex are activated by the sight of another in pain and personal experience of pain (Lamm, Decety, & Singer, 2011). However, as noted above, a critical function of emotional expressions is to convey valence information to guide others behavior and “teach” them the valence of novel objects/actions (Blair, 2003; Fridlund, 1992). Aberrant conscious experience might be important to understand psychopathy, but more fundamental impairments in emotional learning are more likely to be relevant. The insula, amygdala, and striatum are all involved in different aspects of stimulus-reinforcement learning. As such, the systems identified to respond to the pain of others are systems that might associate the valence of the pain to others with the action/object that caused that pain leading to this action/object being valued as aversive. This process is assumed to be impaired in psychopathy (Blair, 1995).

Two psychophysiological studies reported that individuals with psychopathy show reduced autonomic responses to the pain of others (Aniskiewicz, 1979; House & Milligan, 1976). Our own recent fMRI work with youth with psychopathic tendencies has indicated that these youth show reduced neural responses to the sight of another individual’s pain within rostral anterior cingulate cortex, striatum and the amygdala (Marsh et al., *in press*). In short, the response to the pain of others is disrupted in individuals with psychopathic traits.

Processing the disgust of others

Disgusted expressions are also reinforcers, but reinforcers that most frequently provide valence information about foods (Rozin, Haidt, & McCauley, 1993). Disgusted expressions are particularly important for the rapid transmission of taste aversions; the observer is warned not to approach the food that the emoter is displaying the disgust reaction to. FMRI work has shown that the amygdala responds to primary disgust stimuli (i.e. aversive tastes and odors; Small, Gregory, Mak, Gitelman, Mesulam, & Parrish, 2003). Moreover, insula lesions block the acquisition and expression of taste aversion learning (Cubero, Thiele, & Bernstein, 1999). In other words, the insula

allows the representation of the aversive taste that can then be associated with the sensory qualities of the novel food. If disgusted expressions are important for taste aversion learning then they too should recruit the insula. From the earliest studies (e.g. Phillips, Young, Scott, Calder, Andrew, Giampietro, et al., 1998), neuroimaging work shows that they do (see, for a meta-analytic review, Murphy et al., 2003). In addition, neuropsychological work shows that patients with damage to the insula present with selective impairment for the recognition of disgusted expressions (e.g. Calder, Keane, Manes, Antoun, & Young, 2000). In other words, the insula allows the representation of the aversive taste whether this be a primary disgust stimulus (i.e. a taste or odor) or a communication of aversive taste through the disgusted expression of another. This can then be associated with the sensory qualities of the novel food.

With isolated exceptions (Kosson, Suchy, Mayer, & Libby, 2002), the recognition of disgust expressions has not been found to be impaired in individuals with psychopathic tendencies (see, for a meta-analytic review, Marsh & Blair, 2008). Even the Kosson et al. (2002) study only reported impairment when the psychopathic participants responded with their left and not when they responded with their right hands—as such the result cannot be considered to reflect impaired recognition of disgust expressions *per se*. In short, the response to the disgust of others does not appear to be disrupted in individuals with psychopathic traits.

Processing the anger of others

Angry expressions are known to curtail the behavior of others in situations where social rules or expectations have been violated (Averill, 1982). It has been argued that displays of anger or embarrassment do not act as unconditioned stimuli for aversive conditioning or instrumental learning, but rather as important signals to modulate current behavioral responding, particularly in situations involving hierarchy interactions (Blair, 2003). They appear to serve to inform the observer to stop the current behavioral action rather than to convey any information as to whether that action should be initiated in the future. Angry expressions, in short, trigger response reversal (Blair, 2003), a function that inferior frontal cortex is particularly involved in (Cools, Clark, Owen, & Robbins, 2002)—particularly regarding the actual change in response (cf. Budhani, Marsh, Pine, & Blair, 2007). From the earliest work (Blair et al., 1999), neuroimaging studies have shown that the neural response to angry expressions involves inferior frontal cortex (see, for a meta-analytic review, Murphy et al., 2003).

Individuals with psychopathy show no indications of significant impairment in the response to another individual's anger. Impairment in the recognition of anger is not typically reported in this population (see, for a meta-analytic review, Marsh & Blair, 2008). Moreover, the neuroimaging work indicates that the response to angry expressions is intact in youth and adults with psychopathic tendencies (Marsh et al., 2008; Pardini & Phillips, 2010; Passamonti et al., 2010). In short, the response to the anger of others does not appear to be disrupted in individuals with psychopathic traits.

Moral judgment

The past 15 years have seen great change in the understanding the development of morality. Following the relatively long dominance of positions that morality reflected rational thought (Colby & Kohlberg, 1987), positions emerged stressing the importance of emotion, first from data from a clinical population, individuals with psychopathy (Blair, 1995) and then later data with healthy adults (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001). More recently, still there has been a push back against emotion-based positions. This is partly due to a

growing recognition of the complexity of the computations that are subsumed within the term moral.

For instance, it is important to distinguish judgments of “badness,” or affectively negative occurrences (e.g. a hurricane kills five people) from judgments of “immorality” (e.g. a person kills five people; Nichols, 2002). As noted by Nichols, emotion-based systems generate judgments of “badness,” but could not—on their own—generate judgments of “immorality” (an individual killing five people and a hurricane killing five people are both “bad”, but only the first is usually considered as immoral). For an act to be considered “immoral” there must not only be an emotion-based sense of “badness,” but also a representation of the perpetrator’s intent. Indeed, it should be noted that intent information, based on ToM, appears to supersede emotion information during development (Piaget, 1932; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). As such an emotion-based account cannot provide a full explanation of adult moral reasoning. However, such an account can allow an understanding of the basis of moral development (even if it does not allow for an account of the development of all aspects of adult moral reasoning).

According to the IES model, individuals with elevated psychopathic traits show impairment in the amygdala’s role in responding to/learning from fearful, sad, and pain expressions (Blair, 2007). Developmentally, this leads to such individuals regarding care-based transgressions (actions that frighten, upset, or hurt others) as less aversive than healthy individuals do. In contrast, the insula’s role in responding to/learning from disgusted expressions and inferior frontal cortex’s role in responding to angry expressions is considered intact. Consequently, developmentally, it is argued that the processing of disgust-based (actions associated with the disgust of others) and conventional (actions associated with hierarchical violations leading to interpersonal anger) transgressions should be intact. Finally, it is worth also mentioning that the model assumes that the role of vmPFC in representing reinforcement expectancy information is disrupted in psychopathy. This is important to note as reinforcement expectancies guide behavior—both judgments of badness, but also whether actions should be performed or not (Blair, 2007).

In line with this, several studies have shown that individuals with psychopathy show impairment in processing care based transgressions (Blair, 1995, 1997; Koenigs, Kruepke, Zeier, & Newman, 2011). Moreover, the recruitment of the amygdala and vmPFC during care-based reasoning is disrupted in youth and adults with psychopathic tendencies (Glenn, Raine, & Schug, 2008; Harenski, Harenski, Shane, & Kiehl, 2010; Marsh, Finger, Fowler, Jurkowitz, Schechter, Yu, et al., 2011).

Data suggests intact conventional transgressions processing (Blair, 1995, 1997; see also; Nucci & Herman, 1982), as well as disgust-based transgression processing (Glenn, Iyer, Graham, Koleva, & Haidt, 2009) in this population. The Glenn et al. (2009) study is worth describing in more detail. This comprised a very large study with healthy adults ($n=2517$) relating psychopathic traits, as indexed by Levenson’s Self-report Psychopathy Scale (Levenson, Kiehl, & Fitzpatrick, 1995), to performance on the Moral Foundations Questionnaire (Glenn et al., 2009). This is a measure assessing an individual’s commitment to the domains of morality (e.g. the care-based, disgust-based and conventional domains; (Glenn et al., 2009). In line with previous work (Blair, 1995), the Glenn et al. study (2009) reported that higher psychopathic tendencies were related to notably less commitment to care-based norms, but intact commitment to conventional norms. In addition, importantly, this study extended earlier work by also demonstrating: (i) little relationship between psychopathic tendencies and commitment to disgust-based norms; and (ii) that psychopathic tendencies were associated with an increased willingness to violate norms of **any type** for money (Glenn et al., 2009). This last finding is potentially very important. A basic tenet of the IES model is that emotional learning systems allow norms to acquire emotive force, force which guides

attitudes toward these norms. Only one of these emotional learning systems is thought to be disrupted in psychopathic traits, but all are thought to feed reinforcement expectancy information to vmPFC. The vmPFC is thought to represent this information and allow appropriate decision making, particularly in situations where there are multiple reinforcements to be evaluated. In short, the “willingness to violate” data support earlier data indicating impairment in reinforcement-based decision making in individuals with psychopathic tendencies (Blair, Colledge, & Mitchell, 2001a) and critically indicate that this is present whatever the nature of the reinforcement (disgust- and anger-based as well as distress cue-based).

With respect to this latter point, other social reasoning tasks have been shown to rely on the role of vmPFC in representing reinforcement expectancies. For example, in the Prisoner’s Dilemma task, participants are required to choose whether to cooperate with another player or not. Contingencies are set up such that the most reinforcing option is to defect, but only if the other person cooperates. Otherwise, mutual cooperation is the most rewarding option. Healthy participants show greater vmPFC activation when choosing to cooperate presumably because of the greater probability of receiving reward—assuming, as is typically the case, the partner also cooperates (Rilling, Glenn, Jairam, Pagnoni, Goldsmith, Elfenbein, et al., 2007; Rilling, Gutman, Zeh, Pagnoni, Berns, & Kilts, 2002). However, adults with elevated psychopathic traits do not (Rilling et al., 2007). Moreover, adults with elevated psychopathic traits are significantly more likely to defect (Mokros, Menner, Eisenbarth, Alpers, Lange, & Osterheider, 2008; Rilling et al., 2007). The suggestion would be that this is due to their weaker representation of the rewards of cooperation within vmPFC. The same explanation can be used to explain their impaired performance on other social decision making games, such as the Ultimatum and Dictator games, performance similar to that shown by patients with vmPFC damage (Koenigs, Kruepke, & Newman, 2010).

In short, specific aspects of moral judgment are dysfunctional in psychopathy. Learning about the badness of care-based transgressions, which is reliant on the amygdala’s response to distress cue stimuli, is disrupted. However, learning about the badness of disgust-based and conventional transgressions appears intact. Interestingly, although certain types of decision based on this information, particularly choice behavior, requiring the representation of the reinforcements associated with two or more options and necessitating appropriate reinforcement signaling within vmPFC is dysfunctional both in the context of specific forms of moral reasoning task and in other social reasoning tasks more generally.

Conclusions

In conclusion, rather than a unitary mechanism, social cognition and empathy are terms that subsume a variety of computational processes that are mediated by at least partially separable neural systems within the brain. We have distinguished here between cognitive, motor and emotional empathy and even within the category of emotional empathy between systems involved in the response to fear/sadness, pain, disgust and anger. Work with patients with psychopathy is interesting for social cognition because it shows, differently from work with autism, how these computational processes can be selectively disrupted. Unlike patients with autism, individuals with psychopathy have no impairment in cognitive empathy (ToM). Nor do they have impairment in motor empathy. Their impairments are confined to aspects of emotional empathy. But even here the deficits are selective. Responding to and learning from angry and disgusted expressions appears intact. However, learning from and responding to fearful/sad and painful expressions is not. This selective impairment has severe developmental consequences. The individual fails to learn to avoid actions that upset and harm others. Their moral judgment is impaired, but, more importantly,

their behavior toward others is disrupted. The individual becomes more able to commit actions that will harm others if they will lead to advantage for the self.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institute of Mental Health, National Institutes of Health under grant number 1-ZIA-MH00286008. The authors have no conflicts of interest or financial disclosures to report.

Author's contribution to the Work was done as part of the Author's official duties as a NIH employee and is a Work of the United States Government. Therefore, copyright may not be established in the United States.

References

- Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology* 12(2):169–77.
- Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature* 433(7021):68–72.
- Adolphs, R., & Tranel, D. (2004). Impaired judgments of sadness but not happiness following bilateral amygdala damage. *Journal of Cognitive Neuroscience* 16(3):453–62.
- Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Review in Neuroscience* 7(4):268–77.
- Aniskiewicz, A. S. (1979). Autonomic components of vicarious conditioning and psychopathy. *Journal of Clinical Psychology* 35:60–7.
- Averill, J. R. (1982). *Anger and Aggression: An Essay on Emotion*. New York: Springer-Verlag.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition* 21:37–46.
- 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* 42(2):241–51.
- Bastiaansen, J. A., Meffert, H., Hein, S., Huizinga, P., Ketelaars, C., Pijnenborg, M., Bartels, A., Minderaa, R., Keyers, C., & de Bildt, A. (2011). Diagnosing autism spectrum disorders in adults: the use of Autism Diagnostic Observation Schedule (ADOS) module 4. *Journal of Autism Development Disorders* 41(9):1256–66.
- Blair, K. S., Newman, C., Mitchell, D. G., Richell, R. A., Leonard, A., Morton, J., & Blair, R. J. R. (2006). Differentiating among prefrontal substrates in psychopathy: neuropsychological test findings. *Neuropsychology* 20(2):153–65.
- Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57:1–29.
- Blair, R. J. R. (1997). Moral reasoning in the child with psychopathic tendencies. *Personality and Individual Differences* 22:731–9.
- Blair, R. J. R. (1999). Responsiveness to distress cues in the child with psychopathic tendencies. *Personality and Individual Differences* 27:135–45.
- Blair, R. J. R. (2003). Facial expressions, their communicatory functions and neurocognitive substrates. *Philosophical Transactions of the Royal Society, London, B Biological Science* 358(1431):561–72.
- Blair, R. J. R. (2004). The roles of orbital frontal cortex in the modulation of antisocial behavior. *Brain and Cognition* 55(1):198–208.
- Blair, R. J. R. (2005). Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition* 14(4):698–718.

- Blair, R. J. R. (2006). Dissociable systems for empathy. In G. Bock & J. Goode (Eds), *Empathy and Fairness* (pp. 134–41). Chichester: John Wiley and Sons.
- Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Science* 11(9):387–92.
- Blair, R. J. R. (2008). Fine cuts of empathy and the amygdala: dissociable deficits in psychopathy and autism. *Quarterly Journal of Experimental Psychology* 61(1):157–70.
- Blair, R. J. R. (2010). Neuroimaging of psychopathy and antisocial behavior: a targeted review. *Current Psychiatry Reports* 12:76–82.
- Blair, R. J. R., Colledge, E., & Mitchell, D. G. (2001a). Somatic markers and response reversal: is there orbitofrontal cortex dysfunction in boys with psychopathic tendencies? *Journal of Abnormal Childhood Psychology* 29(6):499–511.
- Blair, R. J. R., Colledge, E., Murray, L., & Mitchell, D. G. (2001b). A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies. *Journal of Abnormal Childhood Psychology*, 29(6):491–498.
- Blair, R. J. R., Jones, L., Clark, F., & Smith, M. (1997). The psychopathic individual: A lack of responsiveness to distress cues? *Psychophysiology* 34:192–8.
- Blair, R. J. R., Mitchell, D. G., Richell, R. A., Kelly, S., Leonard, A., Newman, C., & Scott, S. K. (2002). Turning a deaf ear to fear: impaired recognition of vocal affect in psychopathic individuals. *Journal of Abnormal Psychology* 111(4):682–6.
- Blair, R. J. R., Mitchell, D. G. V., & Blair, K. S. (2005). *The Psychopath: Emotion and the Brain*. Oxford: Blackwell.
- Blair, R. J. R., Morris, J. S., Frith, C. D., Perrett, D. I., & Dolan, R. (1999). Dissociable neural responses to facial expressions of sadness and anger. *Brain* 122:883–93.
- Blair, R. J. R., Sellars, C., Strickland, I., Clark, F., Williams, A., Smith, M., & Jones, L. (1996). Theory of mind in the psychopath. *Journal of Forensic Psychiatry* 7:15–25.
- Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience* 2(8):561–7.
- Budhani, S., Marsh, A. A., Pine, D. S., & Blair, R. J. R. (2007). Neural correlates of response reversal: Considering acquisition. *NeuroImage* 34(4):1754–65.
- Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience* 3:1077–8.
- Cardinal, R. N., & Everitt, B. J. (2004). Neural and psychological mechanisms underlying appetitive learning: links to drug addiction. *Current Opinion in Neurobiology* 14(2):156–62.
- Carr, L., Iacoboni, M., Dubeau, M. C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences, USA* 100(9):5497–502.
- Colby, A., & Kohlberg, L. (1987). *The Measurement of Moral Judgement*. New York: Cambridge University Press.
- Cooke, D. J., Michie, C., & Hart, S. (2006). Facets of clinical psychopathy: Toward clearer measurement. In C. J. Patrick (Ed.), *The Handbook of Psychopathy* (pp. 91–106). New York: Guilford Press.
- Cools, R., Clark, L., Owen, A. M., & Robbins, T. W. (2002). Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. *Journal of Neuroscience* 22(11):4563–7.
- Cubero, I., Thiele, T. E., & Bernstein, I. L. (1999). Insular cortex lesions and taste aversion learning: effects of conditioning method and timing of lesion. *Brain Research* 839(2):323–30.
- Dadds, M. R., El Masry, Y., Wimalaweera, S., & Guastella, A. J. (2008). Reduced eye gaze explains “fear blindness” in childhood psychopathic traits. *Journal of the American Academy of Child and Adolescent Psychiatry* 47:455–63.

- Dadds, M. R., Perry, Y., Hawes, D. J., Merz, S., Riddell, A. C., Haines, D. J., Solak, E., & Abeygunawardane, A. I. (2006). Attention to the eyes and fear-recognition deficits in child psychopathy. *British Journal of Psychiatry* 189:280–1.
- Decety, J. (2011). Dissecting the neural mechanisms mediating empathy. *Emotion Review* 3:92–108.
- Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behaviour and Cognitive Neuroscience Review* 3(2):71–100.
- Dolan, M., & Fullam, R. (2004). Theory of mind and mentalizing ability in antisocial personality disorders with and without psychopathy. *Psychology & Medicine* 34(6):1093–102.
- Dolan, M., & Fullam, R. (2006). Face affect recognition deficits in personality-disordered offenders: Association with psychopathy. *Psychology & Medicine* 36(11):1563–9.
- Drevets, W. C., Lowry, T., Gautier, C., Perrett, D. I., & Kupfer, D. J. (2000). Amygdalar blood flow responses to facially expressed sadness. *Biological Psychiatry* 47(8S):160S.
- Fine, C., & Blair, R. J. R. (2000). Mini review: The cognitive and emotional effects of amygdala damage. *Neurocase* 6:435–50.
- Finger, E. C., Marsh, A. A., Blair, K. S., Reid, M. E., Sims, C., Ng, P., Pine, D. S., & Blair, R. J. R. (2011). Disrupted reinforcement signaling in the orbital frontal cortex and caudate in youths with conduct disorder or oppositional defiant disorder and a high level of psychopathic traits. *American Journal of Psychiatry*, 168(2): 834–841.
- Forth, A. E., Kosson, D. S., & Hare, R. D. (2007). *The Psychopathy Checklist: Youth Version*. Toronto: Multi-Health Systems.
- Frick, P. J. (1995). Callous-unemotional traits and conduct problems: a two-factor model of psychopathy in children. *Issues in Criminological and Legal Psychology* 24:47–51.
- Frick, P. J., Bodin, S. D., & Barry, C. T. (2000). Psychopathic traits and conduct problems in community and clinic-referred samples of children: Further development of the psychopathy screening device. *Psychology Assessment* 12(4):382–93.
- Frick, P. J., & Hare, R. D. (2001). *The Antisocial Process Screening Device*. Toronto: Multi-Health Systems.
- Frick, P. J., & Marsee, M. A. (2006). Psychopathy and developmental pathways to antisocial behavior in youth. In: C. J. Patrick (Ed.), *Handbook of Psychopathy* (pp. 353–374). New York: Guilford.
- Fridlund, A. (1992). Darwin's anti-Darwinism in the expression of the emotions in man and animals. *International Review of Emotion* (Vol. 2, pp. 117–37). New York: Wiley.
- Frith, U. (1989). *Autism: Explaining the Enigma*. Oxford: Blackwell.
- Gallagher, M., & Schoenbaum, G. (1999). Functions of the amygdala and related forebrain areas in attention and cognition. *Annals of the New York Academy of Sciences* 877:397–411.
- Glenn, A. L., Iyer, R., Graham, J., Koleva, S., & Haidt, J. (2009). Are all types of morality compromised in psychopathy. *Journal of Personality Disorders* 23:384–98.
- Glenn, A. L., Raine, A., & Schug, R. A. (2008). The neural correlates of moral decision-making in psychopathy. *Molecular Psychiatry* 14:5–6.
- Gordon, H. L., Baird, A. A., & End, A. (2004). Functional differences among those high and low on a trait measure of psychopathy. *Biological Psychiatry* 56(7):516–21.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293: 1971–2.
- Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychology Review* 108(4):814–34.
- Hare, R. D. (2003). *Hare Psychopathy Checklist-Revised (PCL-R)*, 2nd edn. Toronto: Multi Health Systems.
- Harenski, C. L., Harenski, K. A., Shane, M. S., & Kiehl, K. A. (2010). Aberrant neural processing of moral violations in criminal psychopaths. *Journal of Abnormal Psychology* 119(4):863–74.
- Hart, S. D., Forth, A. E., & Hare, R. D. (1990). Performance of criminal psychopaths on selected neuropsychological tests. *Journal of Abnormal Psychology* 99:374–9.

- Hatfield, E., Cacioppo, J. T., & Rapson, R. (1994). *Emotional Contagion*. New York: Cambridge University Press.
- Hoffman, M. L. (1987). The contribution of empathy to justice and moral judgment. In N. Eisenberg & J. Strayer (Eds), *Empathy and its Development* (pp. 47–80). Cambridge: Cambridge University Press.
- House, T. H., & Milligan, W. L. (1976). Autonomic responses to modeled distress in prison psychopaths. *Journal of Personality and Social Psychology* 34:556–60.
- Jackson, P. L., Rainville, P., & Decety, J. (2006). To what extent do we share the pain of others? Insight from the neural bases of pain empathy. *Pain* 125(1–2):5–9.
- Jeon, D., Kim, S., Chetana, M., Jo, D., Ruley, H. E., Lin, S. Y., Rabah, D., Kinet, J. P., & Shin, H. S. (2010). Observational fear learning involves affective pain system and Cav1.2 Ca²⁺ channels in ACC. *Nature Neuroscience* 13(4):482–8.
- Jones, A. P., Happe, F. G., Gilbert, F., Burnett, S., & Viding, E. (2010). Feeling, caring, knowing: different types of empathy deficit in boys with psychopathic tendencies and autism spectrum disorder. *Journal of Child Psychology & Psychiatry* 51(11):1188–97.
- Jones, A. P., Laurens, K. R., Herba, C. M., Barker, G. J., & Viding, E. (2009). Amygdala hypoactivity to fearful faces in boys with conduct problems and callous-unemotional traits. *American Journal of Psychiatry* 166:95–102.
- Karnik, N. S., McMullin, M. A., & Steiner, H. (2006). Disruptive behaviors: conduct and oppositional disorders in adolescents. *Adolescent Medical Clinic* 17(1):97–114.
- Keysers, C., & Gazzola, V. (2006). Toward a unifying neural theory of social cognition. *Progressive Brain Research* 156:379–401.
- Kiehl, K. A. (2006). A cognitive neuroscience perspective on psychopathy: Evidence for paralimbic system dysfunction. *Psychiatry Research* 142:107–28.
- Kimonis, E. R., Frick, P. J., Fazekas, H., & Loney, B. R. (2006). Psychopathy, aggression, and the processing of emotional stimuli in non-referred girls and boys. *Behavioural Sciences & the Law* 24(1):21–37.
- Koenigs, M., Kruepke, M., & Newman, J. P. (2010). Economic decision-making in psychopathy: a comparison with ventromedial prefrontal lesion patients. *Neuropsychologia* 48(7):2198–204.
- Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2011). Utilitarian moral judgment in psychopathy. *Social & Cognitive Affective Neuroscience* 7(6):708–14.
- Kosson, D. S., Suchy, Y., Mayer, A. R., & Libby, J. (2002). Facial affect recognition in criminal psychopaths. *Emotion* 2(4):398–411.
- Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage* 54(3):2492–502.
- LaPierre, D., Braun, C. M. J., & Hodgins, S. (1995). Ventral frontal deficits in psychopathy: Neuropsychological test findings. *Neuropsychologia* 33:139–51.
- LeDoux, J. E. (2007). The amygdala. *Current Biology* 17(20):R868–74.
- Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a non-institutionalized population. *Journal of Personality and Social Psychology* 68:151–8.
- Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., & Baron-Cohen, S. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience* 22(7): 1623–1635.
- Lynam, D. R., Caspi, A., Moffitt, T. E., Loeber, R., & Stouthamer-Loeber, M. (2007). Longitudinal evidence that psychopathy scores in early adolescence predict adult psychopathy. *Journal of Abnormal Psychology* 116(1):155–65.
- Marsh, A. A., & Blair, R. J. R. (2008). Deficits in facial affect recognition among antisocial populations: a meta-analysis. *Neuroscience and Biobehavioral Reviews* 32(3):454–65.
- Marsh, A. A., Finger, E. C., Fowler, K. A., Jurkowitz, I. T., Schechter, J. C., Yu, H. H., Pine, D. S., & Blair, R. J. (2011). Reduced amygdala-orbitofrontal connectivity during moral judgments in youths with disruptive behavior disorders and psychopathic traits. *Psychiatry Research* 194(3):279–86.

- Marsh, A. A., Finger, E. C., Mitchell, D. G. V., Reid, M. E., Sims, C., Kosson, D. S., Towbin, K. E., Leibenluft, E., Pine, D. S., & Blair, R. J. R. (2008). Reduced amygdala response to fearful expressions in children and adolescents with callous-unemotional traits and disruptive behavior disorders. *American Journal of Psychiatry* 165(6):712–20.
- Mineka, S., & Cook, M. (1993). Mechanisms involved in the observational conditioning of fear. *Journal of Experimental Psychology: General* 122:23–38.
- Moffitt, T. E. (1993). The neuropsychology of conduct disorder. *Development and Psychopathology* 5:135–52.
- Mokros, A., Menner, B., Eisenbarth, H., Alpers, G. W., Lange, K. W., & Osterheider, M. (2008). Diminished cooperativeness of psychopaths in a prisoner's dilemma game yields higher rewards. *Journal of Abnormal Psychology* 117(2):406–13.
- Munoz, L. C., & Frick, P. J. (2007). The reliability, stability, and predictive utility of the self-report version of the Antisocial Process Screening Device. *Scandinavian Journal of Psychology* 48:299–312.
- Murphy, F. C., Nimmo-Smith, I., & Lawrence, A. D. (2003). Functional neuroanatomy of emotions: a meta-analysis. *Cognitive, Affective & Behavioral Neuroscience* 3(3):207–33.
- Neumann, C. S., Kosson, D. S., Forth, A. E., & Hare, R. D. (2006). Factor structure of the Hare Psychopathy Checklist: Youth Version (PCL: YV) in incarcerated adolescents. *Psychology Assessment* 18:142–54.
- Nichols, S. (2002). Norms with feeling: toward a psychological account of moral judgment. *Cognition* 84(2):221–36.
- Nucci, L. P., & Herman, S. (1982). Behavioral disordered children's conceptions of moral, conventional, and personal issues. *Journal of Abnormal Child Psychology* 10:411–25.
- Pardini, D. A., & Phillips, M. (2010). Neural responses to emotional and neutral facial expressions in chronically violent men. *Journal of Psychiatry Neuroscience* 35(6):390–8.
- Passamonti, L., Fairchild, G., Goodyer, I. M., Hurford, G., Hagan, C. C., Rowe, J. B., & Calder, A. J. (2010). Neural abnormalities in early-onset and adolescence-onset conduct disorder. *Archives of General Psychiatry* 67(7):729–38.
- Patrick, C. J. (1994). Emotion and psychopathy: Startling new insights. *Psychophysiology* 31:319–30.
- Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., Williams, S. C., Bullmore, E. T., Brammer, M., & Gray, J. A. (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society, London, B Biological Sciences* 265(1408):1809–17.
- Piaget, J. (1932). *The Moral Development of the Child*. London: Routledge and Kegan Paul.
- Preston, S. D., & de Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences* 25(1):1–20.
- Richell, R. A., Mitchell, D. G., Newman, C., Leonard, A., Baron-Cohen, S., & Blair, R. J. (2003). Theory of mind and psychopathy: can psychopathic individuals read the “language of the eyes”? *Neuropsychologia* 41(5):523–6.
- Rilling, J. K., Glenn, A. L., Jairam, M. R., Pagnoni, G., Goldsmith, D. R., Elfenbein, H. A., & Lilienfeld, S. O. (2007). Neural correlates of social cooperation and non-cooperation as a function of psychopathy. *Biological Psychiatry* 61(11):1260–71.
- Rilling, J. K., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., & Kilts, C. (2002). A neural basis for social cooperation. *Neuron* 35(2):395–405.
- Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews in Neuroscience* 2(9):661–70.
- Rozin, P., Haidt, J., & McCauley, C. R. (1993). Disgust. In M. Lewis & J. M. Haviland (Eds), *Handbook of emotions* (pp. 575–94). New York: Guilford Press.
- Saxe, R., & Baron-Cohen, S. (2006). The neuroscience of theory of mind. *Social Neuroscience* 1(3–4), i–ix.
- Schmolck, H., & Squire, L. R. (2001). Impaired perception of facial emotions following bilateral damage to the anterior temporal lobe. *Neuropsychology* 15(1):30–8.

- Seguin, J. R., Boulerice, B., Harden, P. W., Tremblay, R. E., & Pihl, R. O. (1999). Executive functions and physical aggression after controlling for attention deficit hyperactivity disorder, general memory, and IQ. *Journal of Child Psychology & Psychiatry* 40(8):1197–208.
- Shamay-Tsoory, S. G., Harari, H., Aharon-Peretz, J., & Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex* 46(5):668–77.
- Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences* 1156:81–96.
- Small, D. M., Gregory, M. D., Mak, Y. E., Gitelman, D., Mesulam, M. M., & Parrish, T. (2003). Dissociation of neural representation of intensity and affective valuation in human gustation. *Neuron* 39:701–11.
- Stevens, D., Charman, T., & Blair, R. J. R. (2001). Recognition of emotion in facial expressions and vocal tones in children with psychopathic tendencies. *Journal of Genetic Psychology* 162(2):201–11.
- Sylvers, P. D., Brennan, P. A., & Lilienfeld, S. O. (2011). Psychopathic traits and preattentive threat processing in children: a novel test of the fearlessness hypothesis. *Psychology Science* 22(10):1280–7.
- Thioux, M., & Keysers, C. (2010). Empathy: shared circuits and their dysfunctions. *Dialogues in Clinical Neuroscience* 12(4):546–52.
- Widom, C. S. (1976). Interpersonal and personal construct systems in psychopaths. *Journal of Consulting and Clinical Psychology* 44:614–23.
- Widom, C. S. (1978). An empirical classification of female offenders. *Criminal Justice and Behavior* 5:35–52.
- Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences* 107:6753–8.

Two systems for action comprehension in autism: Mirroring and mentalizing

Antonia Hamilton and Lauren Marsh

Introduction

Understanding other minds

Imagine in a café, you order a cup of coffee and soon after, see the barista reaching toward the teabags. You quickly infer that she is about to make tea, but did she mis-hear your order, or is she serving someone else already? The ability to rapidly infer the goal of another person's action and make a guess about her underlying intention is critical in everyday social interaction.

Research on social cognition and the problem of understanding other minds has, over the last 30 years, been largely dominated by the idea of "Theory of mind, that is, the ability to consider the internal, mental states of other individuals. In Premack & Woodruff's (1978) original paper on Theory of Mind, they considered the problem of how to infer another actor's intentions, but research in the 30 years since then has been largely dominated by the question of how to infer an actor's beliefs. This is largely because false-belief tasks provide a clear-cut (and possibly the only) way to assess a participant's representational theory of mind (Dennett, 1978). However, in the last few years, interest has grown in the brain and cognitive systems, which allow us to infer an actor's goal or intention by watching her actions.

The present chapter examines the problem of understanding goals and intentions in other minds, and the integrity of these systems in autism. In the first part, we summarize recent research on action understanding in the typical brain, distinguishing between brain networks associated with mirroring and those associated with mentalizing. In the second part, we examine current theories of action understanding in autism, in relation to recent behavioural and neuroimaging evidence. Finally, we evaluate the data in relation to the theories and consider some important future directions.

Part 1: Two networks in the typical brain

Neuroimaging studies over the last 15 years have identified two distinct brain networks which are reliably engaged when typical individuals engage in non-verbal social interactions including observing actions (and possibly inferring goals), imitating actions, and considering other people's beliefs and desires. These two networks are associated with distinct cognitive functions and theoretical approaches. We briefly review the major and recent studies of each network.

The mirror neuron system

Mirror neurons are defined as single cells which respond when an individual performs an action and observes an equivalent action. Such neurons have been recorded in the premotor and parietal

cortex of the macaque monkey (Fogassi, Ferrari, Gesierich, Rozzi, Chersi, & Rizzolatti, 2005; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). Although individual mirror neurons have not been studied in the same regions in the human brain, neuroimaging evidence suggests that equivalent systems can be found (Van Overwalle, 2009; Caspers, Zilles, Laird, & Eickhoff, 2010). The controversy (Hickok, 2009) over whether the mirror neuron system in monkeys is the same as the system identified in humans has largely been resolved by two recent fMRI studies. The first demonstrated matching fine-scale patterns of activity in parietal cortex during performance and observation of finger and hand actions, which implies that very similar neuronal populations are engaged in each task as predicted by the mirror neuron hypothesis (Oosterhof, Wiggett, Diedrichsen, Tipper, & Downing, 2010). Secondly, Kilner, Neal, Weiskopf, Friston, & Frith (2009) asked participants to alternately perform and observe hand actions during fMRI. Suppression of the BOLD signal in inferior frontal gyrus was found when the action performed matched the previous observed action and when the action observed matched the previous performed action. The best explanation for this pattern of activity is that performed and observed actions both engage the same population of neurons, as required by the mirror neuron hypothesis. Thus, these two studies provide the strongest evidence yet for populations of neurons in the human brain with the same mirror properties as those found in the macaque brain. Throughout this chapter, we use the term “mirror systems” as a compact way to describe the human mirror neuron system without requiring the presence of mirror neurons themselves, and we use the term “mirroring” to refer to activity within classic mirror system regions which is assumed to link representations of performed and observed actions.

Since the discovery of human mirror systems, a number of claims have been made concerning their function. The mirror system seems to match observed actions onto the observer’s own motor system, so it has been claimed that this system allows action comprehension and imitation “from the inside” (Rizzolatti, Giacomo, & Sinigaglia, 2010). Similar mirror processes have been implicated in emotional contagion (Singer, Seymour, O’Doherty, Kaube, Dolan, & Frith, 2004; Wicker, Bruno, Keysers, Plailly, Royet, Gallese, et al., 2003). Some suggest that these processes may provide a fundamental step toward language (Rizzolatti, Giacomo, & Arbib, 1998), empathy (Gallese, 2003a) and even mentalizing (Gallese, Vittorio, & Goldman, 1998) abilities. Thus, the mirror system has been hailed as a unifying basis for social cognition (Gallese, Vittorio, Keysers & Rizzolatti, 2004). However, the evidence for some of these claims remains weak.

In the present section, we focus on the claim that the mirror system provides the brain basis for understanding other people’s actions, goals and intentions. Multiple studies have reported that the core human mirror system regions of inferior parietal lobule (IPL) and premotor cortex are engaged when typical individuals observe another person acting (reviewed in Caspers et al., 2010). But can we go further and consider what cognitive processes might take place in these regions? When we see an action, for example, a child picking an apple, we can represent the action in multiple ways. It is possible to encode the shape of the child’s hand (a kinematic feature), the object the child reaches toward (a goal feature) and the child’s overall intention of picking the apple. The human brain likely represents all these features simultaneously, but can we distinguish how and where these are encoded?

Recent work suggests that kinematic and goal features of observed actions engage slightly different components within the human mirror system. Studies examining kinematic processing in the human brain indicate involvement of both higher order visual systems and inferior frontal gyrus (IFG). For example, if you see a person lift a box, you can normally infer the weight of the box based on kinematic factors such as the velocity of the actor’s lifting action (Hamilton, Joyce, Flanagan, Frith, & Wolpert, 2007). However, this ability is disrupted if repetitive transcranial

magnetic stimulation is used to create a “virtual lesion” (Pascual-Leone, Walsh, & Rothwell, 2000) of the IFG (Pobric & Hamilton, 2006; Hamilton & Grafton, 2006). BOLD responses in IFG are also sensitive to different hand apertures during grasping actions (Hamilton, & Grafton, 2008) and to different grasp types for example, ring pull vs. precision grip (Kilner et al., 2009). Evidence from single cell recordings in macaque monkeys also provides support for the idea that kinematic analysis occurs in area F5 (the monkey homologue of human IFG) as different types of grasp elicit different neuronal firing rates (Bonini, Serventi, Simone, Rozzi, Ferrari, & Fogassi, 2011; Spinks, Kraskov, Brochier, Umiltà, & Lemon, 2008).

In contrast, studies of goal processing suggest that the parietal mirror system, in particular anterior intraparietal sulcus (aIPS), is sensitive to action goals, independent of the kinematics that were used to achieve that goal. Hamilton & Grafton (2006) used a repetition suppression task in which participants watched movies of a hand reaching for a food item or tool during fMRI scanning. Data analysis compared trials where the goal of the action was the same as the previous trial (e.g. take-cookie followed by take-cookie) compared with trials where the goal of the action was different to the previous trial (e.g. take-disk followed by take-cookie). The results show that BOLD signal in just one cortical region, the left aIPS, was suppressed when participants saw a repeated action-goal regardless of the hand trajectory used. This pattern of response is predicted only in brain regions which contain neuronal populations that are sensitive to the manipulated features of the movies (taking a cookie vs. a disk) (Grill-Spector, Henson, & Martin, 2006). This means that aIPS contains neuronal populations which are sensitive to action goals. Oosterhof et al. (2010) also found evidence for the encoding of action goals in aIPS using a multi-voxel pattern analysis method that compared fine-grained activation of voxels across conditions. Further studies found that the IPL also encodes action outcomes, regardless of the action kinematics (Hamilton & Grafton, 2009). In this study the same object was acted upon, only the means by which the goal was achieved was manipulated. Action outcome resulted in differential BOLD responses in the IPL regardless of the action kinematics. Data from monkeys is also compatible with this position, with reports of single neurons which differentiate reach-to-eat and reach-to-place actions in the IPL (Fogassi et al., 2005). Note that goal here is defined very simply in terms of the identity of the object a person grasps, for example, taking a cookie compared with taking a computer disk. More complex action sequences and their goals might be represented elsewhere.

Together, these studies demonstrate that the human mirror system responds selectively to observed actions, and that different types of action processes depend more on different components of the mirror system. In particular, kinematic features of an action are encoded in the frontal mirror system, while goal features are encoded in the parietal mirror system. However, these mirror systems are not necessarily the only brain regions with a role in action understanding. As detailed in the next section, some action comprehension tasks also engage brain areas associated with mentalizing.

The mentalizing system

Mentalizing is the process of attributing mental states (beliefs, desires, and intentions) to another actor. Multiple studies have identified a mentalizing network in the brain, comprising medial prefrontal cortex (mPFC) and temporoparietal junction (TPJ). Temporal poles and precuneus are also sometimes found (see Gallagher & Frith, 2003; Amodio & Frith, 2006; Saxe & Kanwisher, 2003, for reviews). These regions are engaged when reading stories which require mental state attributions (Saxe & Powell, 2006; Young, Dodell-Feder, & Saxe, 2010) or when considering the

beliefs and future actions of others in interactive games (Fletcher et al., 1995). For example, playing rock-paper-scissors encourages participants to think (“he thinks I’ll do rock, but I’ll do scissors and trick him”), and computational models can track this type of belief inference occurring in mPFC and TPJ (Hampton & Bossaerts, 2008; Yoshida, Seymour, Friston, & Dolan, 2010). However, the mentalizing network is not only engaged in tasks requiring explicit verbal belief inference. We focus here on the increasing number of studies that report engagement of this network during non-verbal or minimally verbal tasks in which participants attribute intentions or consider the longer term motivations underlying an action.

One of the earliest non-verbal mentalizing studies recorded brain activity while participants viewed animated triangles moving on the screen (Castelli, Happé, Frith, & Frith, 2000). For some of these animations, typical individuals spontaneously describe the action in terms of the mental states of the triangles (e.g. “the big triangle is coaxing the little triangle”), while for others the action of the triangles is purposeless. Observation of the mentalizing triangles results in activation of mPFC and TPJ, despite the lack of verbal stimuli or instructions.

More recently, spontaneous activation of mentalizing systems during action observation was reported by Brass, Schmitt, Spengler, & Gergely (2007). In this study Brass and colleagues showed participants movies of unusual actions (e.g. turning on a light with your knee). In some cases, the context made the action rational (e.g. turning on a light with your knee because your hands are fully occupied), but in other movies the same action was judged as irrational (turning on the light with your knee when your hands are free). Brass et al. report greater activation in the mentalizing network including TPJ and mPFC when participants viewed irrational actions compared with rational ones. Critically, this activation was not related to the unfamiliarity of the actions because all actions were unusual. Rather, the engagement of TPJ and mPFC reflected the judged rationality of the actions. This study shows that observation of human actions without instructions to mentalize can engage brain regions associated with mentalizing if the observed actions are hard to interpret.

Further studies have refined our knowledge of when action understanding engages mentalizing brain systems. de Lange, Spronk, Willems, Toni, & Bekkering (2008) showed participants images of ordinary actions, actions which had an unusual intention and actions which had unusual kinematic features. This study found that while participants watched actions with an unusual intention, there was greater activity in the STS and mPFC, whereas actions with unusual kinematic features activated the IFG more. This study suggests that both mirror and mentalizing systems are complimentary systems which both contribute to action understanding. The additional recruitment of the mentalizing system for action understanding in social contexts is also reported in a study by Ramsey & Hamilton (2010). In this study, participants watched short movies of a toy animal hiding in one of two locations. Following the hiding phase, an actor came out from behind a curtain, surveyed the possible locations and reached into one to find the toy. Similar to the previously mentioned studies, the results showed complimentary activation of both mirror and mentalizing systems; the IFG was sensitive to action trajectory while the mPFC and right temporal pole were sensitive to successful search behaviour. The design of these studies does not allow strong conclusions about whether participants were attributing beliefs to the actor or only considering intentions, but both studies show that tasks focused on intentions with no explicit belief component are processed differently from tasks that focus on simple goals.

Differential engagement of mentalizing and mirroring systems in the brain can also be driven by task demands. In an fMRI study by Spunt, Satpute, & Lieberman (2011), participants showed

increased BOLD responses in IPL and IFG regions during action observation when participants were asked to think about *how* the actions were being performed. In the same subjects and with the same action stimuli, mPFC and TPJ were more active when participants were asked to think about *why* the actions were being performed. This study shows a nice dissociation between levels of action processing in the brain. It seems that the mirror systems are recruited for kinematic analysis of actions, such as “they are gripping a tin can”, but the mentalizing system is recruited for long-term intentionality judgments, such as “they are recycling the can to save the environment.” Again, this study does not distinguish long-term intentions (“I want to recycle”) from beliefs that underlie the intention (“It is good to recycle”).

Summary

All of these studies suggest that the MNS is not the only brain system engaged in action comprehension, but that more complex tasks and situations may call on the mentalizing network. At least two ways in which mirroring and mentalizing systems might be related can be described (Hamilton, 2008). Under a “mirroring first” model (Figure 21.1, black arrows), full engagement of frontal and parietal mirror regions is a necessary precondition for mentalizing about an observed action. In contrast, in a visual inference model (Csibra & Gergely, 2007), visual information alone is sufficient to determine the goal of an action and engage in mentalizing, and frontal mirror systems are not required. Understanding how the mirroring and mentalizing networks are related is an important area for future research. It is also a critical question in making sense of action understanding in autism. We consider the evidence for the integrity and relationship of mirroring and mentalizing processes in autism in the next section.

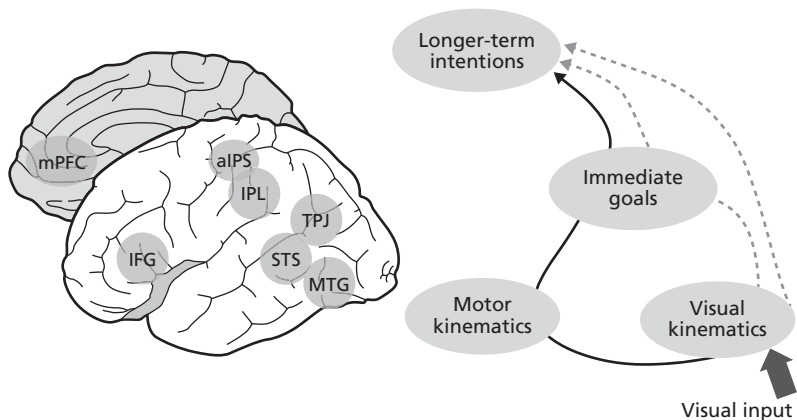


Figure 21.1 Brain and cognitive systems for action comprehension.

Left: Brain systems involved in mirroring (IFG: inferior frontal gyrus; IPL: inferior parietal lobule, aIPS: anterior intra-parietal sulcus), mentalizing (mPFC: medial prefrontal cortex; TPJ: temporoparietal junction), and visual processing of human actions (MTG: middle temporal gyrus; STS: superior temporal sulcus). Right: A sketch of a cognitive model of action processing. Under a mirroring first model (black arrows) visual information processed (MTG/STS) is first matched onto the observers own motor system (IFG), before the goal of the action is extracted (aIPS/IPL) and then longer-term intentions can be defined (TPJ/mPFC). Under a visual inference model (dashed grey arrows), the visual processing (MTG/STS) is sufficient to allow immediate extraction of goals (aIPS/IPL) and longer term intentions (TPJ/mPFC) without the requirement for motor activation.

Part 2: Mirroring and mentalizing in autism

Typically, we automatically attribute goals and intentions to the agents that we observe. However, individuals with autism may not make these same attributions. Currently, there are two competing theories that claim that people with autism have difficulty understanding goals and intentions of others. These are the “mentalizing theory” and the “broken mirror theory.” Each of these theories proposes that one of the two reviewed action understanding networks function atypically in autism. In the mentalizing theory, it is proposed that only the mentalizing network is abnormal, while at least basic processing in the mirror system is normal. In contrast, the broken mirror theory proposes that a core deficit in mirroring leads to difficulties with mentalizing. In the next section, we examine each of these theories, then consider the evidence from each, looking at traditional behavioural tasks, implicit measures, such as eye tracking and EMG, and neuroimaging measures.

Mentalizing theory

There is little disputing the repeated finding that many children and adults with autism have particular difficulties with false belief tasks (Baron-Cohen, Leslie, & Frith, 1985; Frith, 2001). Brain activity in mentalizing regions when participants with autism watch the animated triangles movies is also abnormal (Castelli, Frith, Happé, & Frith, 2002). The mentalizing theory proposes that these difficulties are symptoms of an inability to represent other people’s mental states (Frith, Morton, & Leslie, 1991), or to decouple mental states from reality (Leslie, 1987). Within this field, there is an important distinction between implicit and explicit mentalizing (Apperly & Butterfill, 2009).

Explicit theory of mind is measured with traditional false-belief tasks such as Maxi’s chocolate in which one actor has a false-belief about the location of an object. Participants are typically asked to say or point to the place where Maxi will look for his chocolate (Wimmer & Perner, 1983). Typical children under around 4.5 years old often fail this task, and autistic individuals with a verbal mental age below 9.2 years also tend to fail (Happé, 1995). However, more able individuals with autism often pass false-belief tasks, and may even pass more complex second order tasks (Happé, 1994). Thus, there is a dissociation between the time course of explicit false belief development in typical children (emerging at around 4.5 years and complete by 8 years) and the time course of autism (emerging between 1 and 2 years of age and lasting throughout the lifespan). This has led to a search for precursors to mentalizing and to the investigation of other theories of autism.

In contrast to the late development of explicit mentalizing, implicit mentalizing seems to be present from early infancy (Kovacs, Teglas, & Endress, 2010; Onishi & Baillargeon, 2005) and is measured by recording gaze durations and eye movements when participants view movies in which an actor has a false belief. Recent data demonstrate that even high functioning adults with Asperger’s syndrome who pass verbal false belief tasks fail to show implicit mentalizing in an eye tracking task (Senju, Southgate, White, & Frith, 2009). It is now argued that failure of implicit mentalizing is the core difficulty in autism (Frith, 2012). This resolves the difficulties over the time course of mentalizing failure, because implicit mentalizing develops over the first two years of life at the same time that autism emerges, and implicit mentalizing remains impaired in high-functioning adults with autism. Brain imaging data on implicit mentalizing in autism is not yet available, but it is possible that current tasks such as describing the behaviour of animated triangles tap into implicit mentalizing resources. Brain activation in this task is abnormal in high functioning adults with autism, despite their good explicit theory of mind skills (Castelli et al., 2002).

Research on implicit mentalizing and the precise difference between implicit and explicit tasks is ongoing, and further developments in understanding the role of implicit theory of mind in autism are likely. For present purposes, we contrast a pure mentalizing theory of autism with a

broken mirror theory. The pure mentalizing theory predicts that mentalizing is a single, core deficit in autism and that other social brain systems are unaffected or secondarily affected. For example, basic goal understanding processes should be intact in autism under the mentalizing theory because these do not require the mentalizing network. However, there is still debate over whether difficulties with mentalizing are a single, core deficit in autism or whether these are a consequence of abnormal processing in other social brain systems, for example the mirror system. We consider this question in the next section.

Broken mirror theory

The broken mirror theory claims that developmental failure of the mirror system is the primary social difficulty in autism, and a cause of poor mentalizing. Under this theory, deficits in understanding the kinematic and goal features of an action would lead to further difficulties in understanding emotions and mental states. Initial evidence in support of this theory came primarily from studies of imitation. When typical adults imitate hand actions, the mirror system is activated (Buccino, Binkofski, & Riggio, 2004; Decety, Chaminade, Grèzes, & Meltzoff, 2002; Iacoboni, 1999) and damage to the mirror system in adults causes imitation difficulties (Heilman, Rothi, & Valenstein, 1982). Children with autism may also have trouble with imitation tasks, as summarized in a meta-analysis (Williams, Whiten, & Singh, 2004). Some studies report abnormal brain responses in autistic children during imitation (Dapretto et al., 2006) and action observation (Nishitani, Avikainen, & Hari, 2004; Oberman, Hubbard, McCleery, Altschuler, Ramachandran, & Pineda, 2005). Based on these findings, it was suggested that dysfunction of the mirror system in children with autism might cause first a lack of imitation, and later difficulties in understanding other people's intentions or emotions in social situations (Iacoboni & Dapretto, 2006; Ramachandran & Oberman, 2006; Williams, Whiten, Suddendorf, & Perrett, 2001).

A more recent variant of the broken mirror theory focuses not on comprehension of individual goal directed actions, but on the prediction of actions in a sequence. The account is based on the finding that mirror neurons in parietal cortex encode actions as part of a sequence (Fogassi et al., 2005). For example, some mirror neurons in inferior parietal lobule (IPL) respond selectively when the monkey brings food to his mouth or sees someone bring food to their mouth, but not when bringing a small object toward the shoulder or seeing someone bring an object to their shoulder. They suggest these mirror neurons allow an observer to chain actions together and represent intentions. Building on this work, Cattaneo, Fabbri-Destro, Boria, Pieraccini, Monti, Cossu, et al. (2007) measured electromyographic (EMG) recordings from a jaw-opening muscle (mylohyoid MH) in children when they were performing simple reach-to-eat and reach-to-place actions. In typical children, MH activity increased during the reach phase of a reach-to-eat action, but not of a reach-to-place action, and similar results were found for observation of actions. Thus, typical children chain together the reach and mouth-open actions of an eating sequence, and show similar predictive mouth opening when observing others. In contrast, matched children with autism did not show this anticipatory mouth opening, during either performance or observation. Based on these data, Rizzolatti & Fabbri-Destro (2010) put forward an action-chaining hypothesis of autism. They suggest that predicting actions and inferring intentions in this way is a precursor to mentalizing and belief inference skills. If this is true, then a deficit in action chaining could lead to the social deficits we see in autism (Rizzolatti, Fabbri-Destro, & Cattaneo, 2009).

Contrasting the mentalizing and broken mirror theories, some important differences emerge. The traditional mentalizing theory derives from a symbolic, abstract view of cognition (Leslie, 1987), while the broken mirror account is associated with an embodied approach which emphasizes the role of simulation in understanding others (Gallese, 2003b; Goldman, 2006). Similarly, the

mentalizing theory places the primary deficit in “high level” reasoning about and representation of mental states, and assumes that abnormal social behaviour in simple situations are a consequence of this. Meanwhile, the broken mirror theory focuses on lower level problems with imitation and assumes that failure on theory of mind tasks arises because simpler simulation mechanisms are dysfunctional in autism. Neither theory attempts to account for all the characteristics of autism, including non-social problems such as repetitive behaviours or differences in perceptual processing that might be attributed to weak central coherence (Frith & Happé, 1994).

To test and discriminate between the mentalizing theory and the broken mirror theory, it is interesting to examine the realms where they overlap. In particular, goals and intentions are relevant to both theories. Mirror neurons in macaque monkeys respond only to goal-directed actions (Fogassi et al., 2005; Gallese et al., 1996; Umiltà, Kohler, Gallese, Fogassi, Fadiga, Keysers, et al., 2001), so goals are key to the original idea of mirror neuron function. The human mirror system seems to be more general, with some response even to actions without a goal, but goal-directed actions are a powerful stimulus which robustly activate this system (Gazzola, Rizzolatti, Wicker, & Keysers, 2007; Iacoboni, Molnar-Szakacs, Gallese, Buccino, Mazziotta, & Rizzolatti, 2005; Koski, Wohlschläger, Bekkering, Woods, Dubeau, Mazziotta, et al., 2002). Damage to the human parietal mirror system, e.g. from stroke, is known to cause difficulties with understanding and performing meaningful or goal-directed actions (Buxbaum, Kyle, & Menon, 2005). Therefore, a lack of goal understanding in autism is a key prediction of the broken mirror theory.

In this section, we evaluate the claims that either the whole mirror system or the ability to chain actions in a sequence is abnormal in autism. We focus mainly on recent studies that use implicit (eyetracking or EMG) measures of action comprehension and on neuroimaging studies. A large number of studies of imitation in autism have been reviewed in greater depth elsewhere (Hamilton, 2008; Southgate & Hamilton, 2008; Williams et al., 2001).

Behavioural studies of action understanding in autism

Multiple studies have reported poorer imitation performance in children with autism compared with typical children on general batteries of imitation tasks, including imitation of meaningless actions, mimicry of facial expressions and the spatial perspective taking component of imitation. These results have led to the claim that there is a global imitation impairment in autism (Williams et al., 2004). However, more recent studies suggest autistic children successfully imitate when explicitly instructed to do so, whether imitating hand actions (Beadle-Brown, 2004) or facial expressions (McIntosh, Reichmann-Decker, Winkielman, & Wilbarger, 2006). They also show better performance in a highly structured imitation task than in a task requiring spontaneous imitation (Hepburn & Stone, 2006).

An interesting comparison in imitation studies is between imitation of a goal and imitation of kinematic features or action style, because these fall at different levels of the action hierarchy. Hobson and colleagues (Hobson & Hobson, 2008; Hobson & Lee, 1999) tested children with autism on a novel action imitation task. For example, children were shown how to scrape two objects together to make a sound and were asked to copy. Children with autism were able to perform the same, goal directed action, but failed to mimic the style (loud or soft) with which the action was performed. Intact goal-directed imitation in children with autism has also been seen in a simple hand movement task. Autistic children and controls matched for verbal mental age were tested on Bekkering’s goal directed imitation task (Bekkering, Wohlschläger, & Gattis, 2000). In this task children were asked to copy an experimenter who touched one of two targets on the table in front of them. The experimenter sometimes made an ipsilateral movement of her hand to the nearest dot (e.g. left hand to left dot) and sometimes made a contralateral movement of her hand to the

further dot (e.g. right hand to left dot). Both groups of children accurately imitated the action goal, i.e. they touched the appropriate dot on the table. More importantly, both typical and autistic children made systematic hand errors; when the demonstrator moved her hand across her body, the child correctly imitated the goal, but failed to use the appropriate hand (Hamilton, Brindley, & Frith, 2007). This is the pattern of behaviour taken by Bekkering and colleagues to be a signature of goal directed imitation. Children with autism are not imitating only the outcome of the action, but must be identifying the goal and selecting how to achieve that goal. Thus, the data provides evidence that both typical and autistic children have a goal hierarchy and can understand and imitate the goal of an adult's action. Furthermore, children with autism can and go beyond the immediately visible goal of an adult's action and imitate goals which they had not seen achieved. Two independent studies (Aldridge, Stone, Sweeney, & Bower, 2000; Carpenter, Pennington, & Rogers, 2001) found that children with autism completed the action of pulling apart the dumb-bell even when the adult demonstrator had never successfully performed the action. In summary, it seems that autistic children are able to imitate actions, when given clear and explicit instructions to do so. The behavioural evidence reviewed here suggests that simple goal representation is intact in autism, contrary to the predictions of the broken mirror hypothesis.

Understanding of more complex goals or action sequences is being increasingly studied in autism, but results are contradictory. One study using a picture ordering task to compare understanding of mental state sequences to simpler goal-directed action sequences found that individuals with autism had no problems understanding and ordering the goal directed sequences (Baron-Cohen, Leslie, & Frith, 1986). However, a similar study found participants with autism did have trouble understanding object-directed action sequences (Zalla, Labruyere, & Georgieff, 2006), but surprisingly not interactive action sequences.

More recently, a study by Boria, Fabbri-Destro, Cattaneo, Sparaci, Sinigaglia, Santelli, et al. (2009) demonstrated poorer understanding of subsequent actions in children with autism. In this study, children were shown static images of a hand either touching an object, grasping-to-use it or grasping-to-place it. Children were asked what the actor was doing and why. Children with autism were able to distinguish touching and grasping actions. They were also able to identify subsequent use of the object, as well as typically developing children in the grasp-to-use condition. However, their performance was substantially poorer when identifying the grasp-to-place actions, with object-use dominating their responses, despite the grasp type rendering this action implausible. Boria and colleagues argue that children with autism are unable to use the motor information to make an inference about the subsequent action, providing evidence for the action chaining theory. However, in their second similar experiment, children with autism were able to identify grasp-to-place actions if an image of the end goal was also present. Boria argues that this evidence corroborates their initial finding and children with autism are not just making stereotyped, object-use responses. An alternative explanation for this improved ability in the second experiment could be that the imagination demands are reduced as the action end point is visible. A better test of this effect should test different, dynamic grasps with the possible end points visible. This will reduce the imagination demand of the task and will require correct analysis of the motor properties of the grasp to infer the subsequent action.

Implicit measures of action understanding in autism

Eye tracking studies of action observation have also been used to assess mirror neuron function in autistic children. Typically, eye movements during action observation and action execution are predictive of the actions that they are monitoring. It has been suggested that these predictive eye movements are reflective of mirror neuron function as eye movements during action

observation mirror those during action execution (Flanagan & Johansson, 2003). In support of this claim (Cannon & Woodward, 2008) demonstrated that predictive eye movements during action observation are disrupted by simultaneous performance of sequential finger movements, but not by the rehearsal of sequences of numbers. In a study of autistic 5-year olds (Falck-Ytter, 2010) demonstrated that infants with autism were able to anticipate actions to the same degree as typical infants and adults. This finding suggests that even young children with autism are able to predict the actions of others and provides evidence against impaired action chaining in autism.

However, other studies of action chaining in autism do suggest difficulties. Cattaneo et al. (2007), as described earlier, showed that children with autism failed to produce predictive MH muscle activation during the performance or observation of a reach-to-eat action, in contrast to typical control children. They argue that this indicates a failure of action chaining in participants with autism. One limitation in this study is the failure to exclude dyspraxia in the autistic sample of participants; dyspraxia is often comorbid with autism (Ming, Brimacombe, & Wagner, 2007) and impacts on motor control, but it is not linked to mentalizing.

Further evidence for impaired action chaining in autism comes from a study by Fabbri-Destro, Cattaneo, Boria, & Rizzolatti (2009) who used a similar methodology to that of Johnson-Frey, McCarty, & Keen (2004). In this study, children with and without autism were asked to pick up a block and move it to either a small or large container whilst their movement time was measured. Throughout the experiment, the task demands of the reach action remained constant. However, manipulating the size of the container increased the task demands of the place action. Despite the controlled demands of the reach action across conditions, typically developing children modified the speed of the initial reach action such that they were slower when the following action was harder and faster when the following action was easier. This bias is thought to reflect future planning of the second action in the sequence. In children with autism, the speed of the reach action was not biased by the difficulty of the following action, indicating a lack of action planning. Overall, the evidence for impaired action chaining in autism is mixed. Eye-tracking studies show that online action prediction is functioning typically in autistic children. Studies that use more complex action sequences do reveal differences between typical and autistic children, although they fail to control for motor ability in their tasks. Further research is needed to assess the action chaining account of the broken mirror hypothesis.

Neuroimaging studies of action understanding in autism

Neuroimaging techniques provide the most rigorous tests of the integrity of the mirror system in autism. A number of early studies report differences between typical and autistic participants. For example, Oberman et al. (2005) report reduced mu wave suppression during observation and execution of hand actions in typical participants, but mu suppression only occurred during execution tasks in the autistic participants. In addition, Théoret, Halligan, Kobayashi, Fregni, Tager-Flusberg, & Pascual-Leone (2005) demonstrated that motor evoked potentials, induced by transcranial magnetic stimulation during action observation were reduced for autistic participants. However, no group differences in magneto-encephalographic recordings were found between typical and autistic participants during the observation of hand actions (Avikainen, Kulomäki, & Hari, 1999). It is important to note that all of these studies used measures with very limited localization of effects and participant numbers were low.

fMRI studies provide evidence with better spatial resolution and can identify specific brain abnormalities in a more convincing way. Dapretto, Davies, Pfeifer, Scott, Sigman, Bookheimer, et al. (2006) conducted the first study to provide evidence for the broken mirror hypothesis with fMRI. In their study, participants were asked to observe and imitate emotional facial

expressions during fMRI scanning. They report reduced activation in the IFG component of the mirror system during observation and imitation in autistic participants. Furthermore, the amount of activation significantly correlated with autistic symptom severity. However, imitation of emotional facial expressions is not a goal-directed action task and it is very different from the original hand-grasping studies that were used to study the mirror neuron system in monkeys (Gallese et al., 1996). Therefore, this study provides only weak evidence for the broken mirror hypothesis.

In a more comparable study of hand actions, Dinstein, Thomas, Humphreys, Minshew, Behrmann, & Heeger (2010) asked participants to perform and observe sequences of simple hand postures during fMRI scanning. They report no group differences between autistic and typical participants during observation or execution of hand postures in mirror neuron regions. In addition, autistic participants demonstrated normal movement selectivity for repeated hand postures in left anterior intraparietal sulcus (aIPS) and ventral premotor cortex (vPM) in both observation and execution conditions. This study provides the first robust evidence against mirror system dysfunction in autism.

Only one study has tried to assess the integrity of both mirror and mentalizing systems in autism in the same study (Marsh & Hamilton, 2011). Manipulation of action rationality was used as a tool to engage the mentalizing system. As previously reported, Brass et al. (2007) demonstrated that irrational actions automatically activate the mentalizing system in the typical observer, even with no prior instruction to mentalize. By using matched rational and irrational action stimuli Marsh and Hamilton (2011) were able to dissociate mirroring and mentalizing systems in the autistic brain in a non-verbal, action observation task.

Eighteen adults with autism and 19 age and IQ-matched typical adults completed the experiment. They watched movies of simple, goal-directed reach actions to either a piece of food or a tool during fMRI scanning. Some actions were rational (Figure 21.2, R1 & R2) while in others the hand took an irrational route to reach the target object (Figure 21.2, I1 & I2). Control movies depicting a shape drifting across the screen were also shown. The results showed that both typical and autistic participants engage mirror regions, in particular left aIPS when observing hand actions. In addition, this area was also sensitive to action goals in both participant groups. As the left aIPS is the established goal processing region of the mirror system as defined in Hamilton & Grafton (2006, 2008), this result provides evidence against a global mirror neuron deficit in autism and corroborates behavioural evidence that suggests that goal understanding is intact in autism.

In contrast, differences between the typical and autistic participants emerged when regions outside the mirror system were examined, and when action rationality was considered. In both typical and autistic participants, the right aIPS was activated for irrational actions compared with rational actions. However, in the mPFC, only typical participants differentiate irrational from rational actions. mPFC activity in the autistic participants remained the same regardless of the rationality of the observed action. These results demonstrate that, within the same group of participants, responses in the mirror system to observed actions can be normal while responses in the mentalizing system are abnormal.

Summary

Evidence for the integrity of mirroring and mentalizing brain systems in autism has been reviewed above. In typical individuals, the mirror system encodes action kinematics and goals, while the mentalizing system plays a role in making inferences about the actors' beliefs and intentions. Evidence for poor mentalizing in autism is clear cut, but there is much less support for the proposal that this social difficulty originates in failure of mirror systems. Many studies have demonstrated

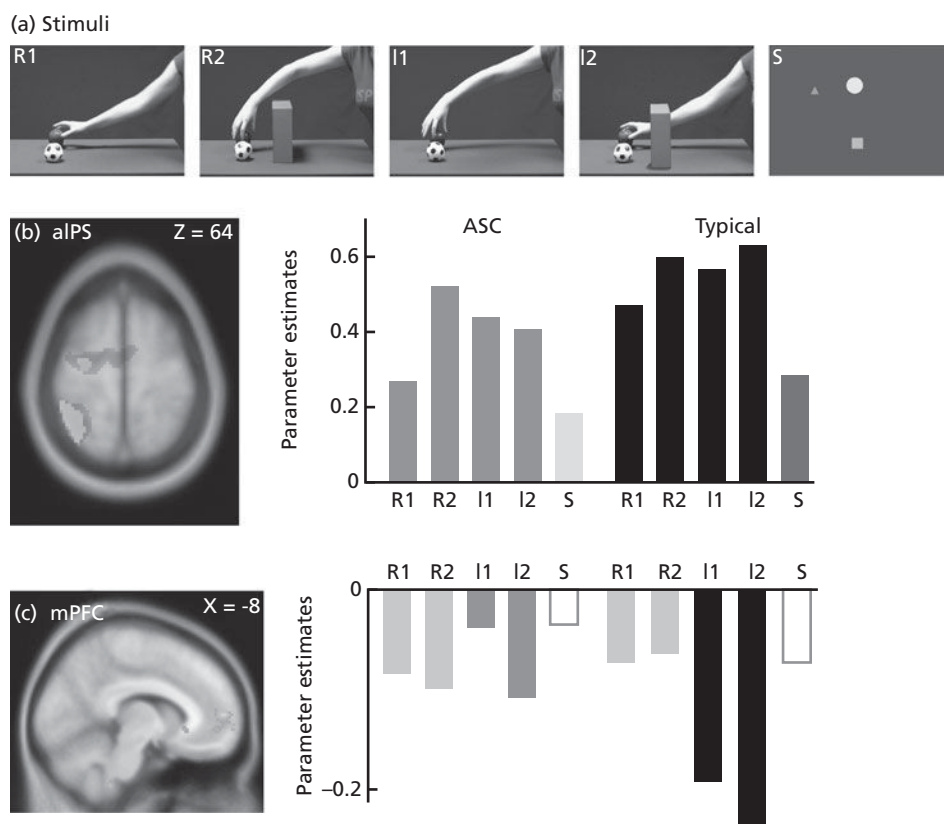


Figure 21.2 Responses of mirroring and mentalizing brain systems in autism. (a) Still frames of the five movie types used in Marsh & Hamilton (2011). In each movie the hand started on the right, moved across to pick up an object and returned its original position. R1: rational action, R2: rational action with a barrier, I1: irrational action, I2: irrational action with a barrier, S: control movie showing three shapes, one of which moved linearly across the screen. (b) Activity in left alPS was greater during the observation of hand actions compared with moving shapes in both autism and typical participants. (c) Activity in mPFC was sensitive to action rationality in the typical group, but not in the autism group.

good goal understanding in autism, together with normal brain responses in mirror systems. However, people with autism may have difficulty understanding sequences of actions, or chaining actions together and this area warrants further exploration.

Conclusions

From the studies reviewed in this chapter, no clear cut evidence emerges for a fundamental mirroring system deficit in autism. Behavioural studies have shown that people with autism have a good understanding of action goals. Furthermore, two independent neuroimaging studies have reported that the parietal component of the mirror system is functioning typically in individuals with autism. Some evidence for the action chaining account exists, but stringent neuroimaging studies need to test this further. Few studies have directly tested the integrity of mentalizing

systems in relation to action understanding in autism, but initial reports suggest that this may be functioning atypically.

An important future direction in this field is to establish the relationship between the mirror system and the mentalizing system. How does kinematic and goal information about actions translate into an understanding of intention? Action rationality is a new tool that can tap in to both mirror and mentalizing systems and studies comparing rational and irrational actions may be able to provide us with a better understanding of the interactions between mirroring and mentalizing. However, a better understanding of what action rationality is and why irrational actions engage the mentalizing system is also needed. Implicit measures, such as eye-tracking, give us insight into the fast, automatic processing of actions and can allude to subtle differences in perception in autism.

References

- Aldridge, M. A., Stone, K. R., Sweeney, M. H., & Bower, T. G. R. (2000). Preverbal children with autism understand the intentions of others. *Developmental Science* 3:294–301.
- Amodio, D., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews of Neuroscience* 7(4):268–77.
- Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review* 116(4):953–70.
- Avikainen, S., Kulomäki, T., & Hari, R. (1999). Normal movement reading in Asperger subjects. *Neuroreport* 10(17):3467–70. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10619627>
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind.” *Cognition* 21:37–46.
- Baron-Cohen, S., Leslie, A., & Frith, U. (1986). Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology* 4:113–25.
- Beadle-Brown, J. (2004). Elicited imitation in children and adults with autism : the effect of different types of actions. *Journal of Applied Research in Intellectual Disabilities* (1991), 17:37–48.
- Bekkering, H., Wohlschlaeger, A., & Gattis, M. (2000). Imitation of Gestures in Children is Goal-directed. *Quarterly Journal of Experimental Psychology* 53(1):153–64.
- Bonini, L., Serventi, F. U., Simone, L., Rozzi, S., Ferrari, P. F., & Fogassi, L. (2011). Grasping neurons of monkey parietal and premotor cortices encode action goals at distinct levels of abstraction during complex action sequences. *Journal of neuroscience : Official Journal of the Society for Neuroscience* 31(15):5876–86.
- Boria, S., Fabbri-Destro, M., Cattaneo, L., Sparaci, L., Sinigaglia, C., Santelli, E., Cossu, G., & Rizzolatti, G. (2009). Intention understanding in autism. *PloS one* 4(5):e5596.
- Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating action understanding: inferential processes vs. action simulation. *Current Biology* 17:2117–121.
- Buccino, G., Binkofski, F., & Riggio, L. (2004). The mirror neuron system and action recognition. *Brain and Language* 89:370–6.
- Buxbaum, L. J., Kyle, K. M., & Menon, R. (2005). On beyond mirror neurons: internal representations subserving imitation and recognition of skilled object-related actions in humans. *Brain Research. Cognitive Brain Research* 25(1):226–39.
- Cannon, E. N., & Woodward, A. L. (2008). Action anticipation and interference : A test of prospective gaze, Paper presented at the 30th Annual Conference of the Cognitive Science Society.
- Carpenter, M., Pennington, B. F., & Rogers, S. J. (2001). Understanding of others’ intentions in children with autism. *Journal of Autism and Developmental Disorders* 31(6):589–99. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/11814270>
- Caspers, S., Zilles, K., Laird, A. R., & Eickhoff, S. B. (2010). ALE meta-analysis of action observation and imitation in the human brain. *NeuroImage* 50(3):1148–67.

- Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain: Journal of Neurology* 125(Pt 8):1839–49. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12135974>
- Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage* 12(3):314–25.
- Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., & Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of the National Academy of Sciences, USA* 104(45):17825–30.
- Csibra, G., & Gergely, G. (2007). Obsessed with goals: Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica* 124(1):60–78.
- Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., & Iacoboni, M. (2006). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* 9(1):28–30.
- de Lange, F. P., Spronk, M., Willems, R. M., Toni, I., & Bekkering, H. (2008). Complementary systems for understanding action intentions. *Current biology: CB* 18(6):454–7.
- Decety, J., Chaminade, T., Grèzes, J., & Meltzoff, A. N. (2002). A PET exploration of the neural mechanisms involved in reciprocal imitation. *NeuroImage* 15(1): 265–72.
- Dennett, D. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences* 4: 568–70.
- di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research* 91: 176–80.
- Dinstein, I., Thomas, C., Humphreys, K., Minshew, N., Behrmann, M., & Heeger, D. J. (2010). Normal movement selectivity in autism. *Neuron* 66(3):461–9.
- Fabbri-Destro, M., Cattaneo, L., Boria, S., & Rizzolatti, G. (2009). Planning actions in autism. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation cérébrale* 192(3):521–5.
- Falck-Ytter, T. (2010). Young children with autism spectrum disorder use predictive eye movements in action observation. *Biology Letters* 6(3):375–8.
- Flanagan, R. J., & Johansson, R. (2003). Action plans used in action observation. *Nature* 424:769–71.
- Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of “theory of mind” in story comprehension. *Cognition* 57: 109–28.
- Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308(5722):662–7.
- Frith, U. (2001). Mind blindness and the brain in autism. *Neuron* 32(6):969–79. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/11754830>
- Frith, U. (2012). Why we need cognitive explanations of autism, 38 Bartlett Lecture 2010. *Quarterly Journal of Experimental Psychology (Hove)* 65(11):2073–92.
- Frith, U., & Happé, F. (1994). Autism: beyond “theory of mind.” *Cognition* 50(1–3):115–32. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12212920>
- Frith, U., Morton, J., & Leslie, A. (1991). The cognitive basis of a biological disorder: autism. *Trends in Neurosciences* 14(10):433–8.
- Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of “theory of mind.” *Trends in Cognitive Sciences* 7(2):77–83. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12584026>
- Gallese, V. (2003a). The roots of empathy: the shared manifold hypothesis and the neural basis of intersubjectivity. *Psychopathology* 36(4):171–80.
- Gallese, V. (2003b). The manifold nature of interpersonal relations: the quest for a common mechanism. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences* 358(1431):517–28.
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119:593–609.

- Gallese, V., & Goldman, A. (1998). Mirror neurons and the mind-reading. *Trends in Cognitive Sciences* 2(12):493–501.
- Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences* 8(9):396–403.
- Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *NeuroImage* 35(4):1674–84.
- Goldman, A. (2006). *Simulating Minds*. Oxford: Oxford University Press.
- Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences* 10(1):14–23.
- Hamilton, A. F. de C., Brindley, R. M., & Frith, U. (2007). Imitation and action understanding in autistic spectrum disorders: How valid is the hypothesis of a deficit in the mirror neuron system? *Neuropsychologia* 45(8):1859–68.
- Hamilton, A. F. de C., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. *Journal of Neuroscience* 26:1133–7.
- Hamilton, A. F. de C., & Grafton, S. T. (2008). Action outcomes are represented in human inferior frontoparietal cortex. *Cerebral Cortex* 18(5):1160–8.
- Hamilton, A. F. de C., & Grafton, S. T. (2009). Repetition suppression for performed hand gestures revealed by fMRI. *Human Brain Mapping* 30(9):2898–906. doi:10.1002/hbm.20717
- Hamilton, A. F. de C., Joyce, D. W., Flanagan, J. R., Frith, C. D., & Wolpert, D. M. (2007). Kinematic cues in perceptual weight judgement and their origins in box lifting. *Psychological Research* 71(1):13–21.
- Hampton, A. N., & Bossaerts, P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Signals* 105(18):6741–6.
- Happé, F. G. (1994). An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders* 24(2):129–54. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/8040158>
- Happé, F. G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development* 66(3):843–55. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/7789204>
- Heilman, K. M., Rothi, L. J., & Valenstein, E. (1982). Two forms of ideomotor apraxia. *Neurology* 32(4):342–6.
- Hepburn, S. L., & Stone, W. L. (2006). Longitudinal research on motor imitation in autism. *Imitation and the Social Mind*. New York: Guildford Press.
- Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience* 21(7):1229–43.
- Hobson, R. P., & Hobson, J. A. (2008). Dissociable aspects of imitation: a study in autism. *Journal of Experimental Child Psychology* 101(3):170–85.
- Hobson, R. P., & Lee, A. (1999). Imitation and identification in autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 40(4):649–59. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10357170>
- Iacoboni, M. (1999). Cortical mechanisms of human imitation. *Science* 286(5449):2526–8.
- Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews of Neuroscience* 7(12):942–51.
- Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology* 3(3):e79.
- Johnson-Frey, S., McCarty, M., & Keen, R. (2004). Reaching beyond spatial perception: Effects of intended future actions on visually guided prehension. *Visual Cognition* 11(2):371–99.
- Kilner, J. M., Neal, A., Weiskopf, N., Friston, K. J., & Frith, C. D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. *Journal of Neuroscience* 29(32):10153–9.

- Koski, L., Wohlschläger, A., Bekkering, H., Woods, R. P., Dubeau, M.-C., Mazziotta, J. C., & Iacoboni, M. (2002). Modulation of motor and premotor activity during imitation of target-directed actions. *Cerebral Cortex* 12(8):847–55. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12122033>
- Kovacs, A. M., Teglas, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science* 330(6012):1830–4.
- Leslie, A. M. (1987). Pretense and representation : The origins of “theory of mind.” *Cognitive Development* 94(4):412–26.
- Marsh, L., & Hamilton, A. (2011). Dissociation of mirroring and mentalizing systems in autism. *NeuroImage* 56(3):1511–19.
- McIntosh, D. N., Reichmann-Decker, A., Winkielman, P., & Wilbarger, J. L. (2006). When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. *Developmental Science* 9(3):295–302.
- Ming, X., Brimacombe, M., & Wagner, G. C. (2007). Prevalence of motor impairment in autism spectrum disorders. *Brain & Development* 29(9):565–70.
- Nishitani, N., Avikainen, S., & Hari, R. (2004). Abnormal imitation-related cortical activation sequences in Asperger's syndrome. *Annals of Neurology* 55(4):558–62.
- Oberman, L. M., Hubbard, E. M., McCleery, J. P., Altschuler, E. L., Ramachandran, V. S., & Pineda, J. A. (2005). EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Brain Research. Cognitive Brain Research* 24(2):190–8.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science* 308(5719):255–8.
- Oosterhof, N. N., Wiggett, A. J., Diedrichsen, J., Tipper, S. P., & Downing, P. E. (2010). Surface-based information mapping reveals crossmodal vision-action representations in human parietal and occipitotemporal cortex. *Journal of Neurophysiology* 104(2):1077–89.
- Pascual-Leone, A., Walsh, V., & Rothwell, J. (2000). Transcranial magnetic stimulation in cognitive neuroscience—virtual lesion, chronometry, and functional connectivity. *Current Opinion in Neurobiology* 10(2):232–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10753803>
- Pobric, G., & Hamilton, A. F. DeC. (2006). Action understanding requires the left inferior frontal cortex. *Current Biology* 16:424–9.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 4:515–26.
- Ramachandran, B. V. S., & Oberman, L. M. (2006). Broken mirrors: A theory of autism. *Scientific American* 295(5):62–9.
- Ramsey, R., & Hamilton, A. F. DeC. (2010). How does your own knowledge influence the perception of another person's action in the human brain? *Social Cognitive and Affective Neuroscience* 7(2):242–51.
- Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neuroscience* 21(5):188–94.
- Rizzolatti, G., & Fabbri-Destro, M. (2010). Mirror neurons: from discovery to autism. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 200(3–4), 223–37.
- Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature Reviews Neuroscience* 11(4):264–74.
- Rizzolatti, G., Fabbri-Destro, M., & Cattaneo, L. (2009). Mirror neurons and their clinical relevance. *Neurology* 5(1):24–34
- Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science* 17(8): 692–9.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people The role of the temporo-parietal junction in “theory of mind.” *NeuroImage* 19:1835–42.
- Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science* 325(5942):883–5.

- 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* 303:1157–62.
- Southgate, V., & Hamilton, A. F. De C. (2008). Unbroken mirrors: Challenging a theory of autism. *Trends in Cognitive Sciences* 12(6):225–9.
- Spinks, R. L., Kraskov, A., Brochier, T., Umiltà, M. A., & Lemon, R. N. (2008). Selectivity for grasp in local field potential and single neuron activity recorded simultaneously from M 1 and F 5 in the awake macaque monkey. *Journal of Neuroscience : Official Journal of the Society for Neuroscience* 28(43):10961–71.
- Spunt, R. P., Satpute, A. B., & Lieberman, M. D. (2011). Identifying the what, why, and how of an observed action: an fMRI study of mentalizing and mechanizing during action observation. *Journal of Cognitive Neuroscience* 23(1):63–74.
- Théoret, H., Halligan, E., Kobayashi, M., Fregni, F., Tager-Flusberg, H., & Pascual-Leone, a. (2005). Impaired motor facilitation during action observation in individuals with autism spectrum disorder. *Current Biology CB* 15(3):R 84–5.
- Umiltà, M. a, Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I know what you are doing. a neurophysiological study. *Neuron* 31(1):155–65. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/11498058>
- Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping* 30(3):829–58.
- Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: The common neural basis of seeing and feeling disgust. *Neuron* 40(3):655–64. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/14642287>
- Williams, J. H. G., Whiten, A., & Singh, T. (2004). A systematic review of action imitation in autistic spectrum disorder. *Journal of Autism and Developmental Disorders* 34(3):285–99. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15264497>
- Williams, J. H., Whiten, A., Suddendorf, T., & Perrett, D. I. (2001). Imitation, mirror neurons and autism. *Neuroscience Biobehavioral Review* 25(4):287–95.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13:103–28.
- Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *Journal of Neuroscience* 30(32):10744–51.
- Young, L., Dodell-Feder, D., & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia* 48(9):2658–64.
- Zalla, T., Labruyere, N., & Georgieff, N. (2006). Goal-directed action representation in autism. *Journal of Autism Development and Disorders* 36(4):527–40.

Autism: Self and others

Peter R. Hobson and Jessica A. Hobson

Introduction

For some scientists, the self is a fiction. Yet how I think of myself as a thinker, how I experience my potential to choose and to act, how I assume responsibility and have feelings of pride, guilt or shame over the ways I behave, how I am subject to states of admiration, envy, or jealousy in relation to others, or how I think and feel about my qualities as a person extending from the past through the present to the future—indeed, how frequently I use the word “I” in communicating with other people, to express my self-anchored perspective—all these qualities of experience testify to the significance of a sense of self and the importance of “I-thoughts.”

If being a self is so central a part of human mental life, then we might do well to consider the varieties of experience that contribute to selfhood, and to investigate how these evolve over the course of development. The study of atypical development—and in the present instance, the study of autism—might add something additional in allowing us to distinguish among potentially separable components of the self, only some of which are compromised in cases of disorder. We might also be in a position to trace distinct developmental pathways to fully-fledged self-experience.

The story of self-development is one in which an individual is both connected to, and differentiated from, the non-personal world on the one hand and people on the other. Therefore we need to consider the kinds of relation that exist between persons with autism and the world, and build up a picture of the kinds of connectedness and differentiation that they display and experience with people and things (Hobson, 1990). For instance, some but not all of our self-experience appears to be tied in with experience of other embodied people who are recognized to be selves in their own right, and whose subjective life may include their having feelings and thoughts toward ourselves. To the extent that self-other relations, and more specifically experiences of other people's attitudes to the self, are atypical among young children with autism, this might have profound implications for the structure and content of their developing self-awareness.

It is sometimes helpful to distinguish between unreflective and reflective levels of self-awareness. Expressions of a person's senses of (unreflected-upon) self include the ability to walk around furniture in a room, or (arguably) to maintain appropriate distance from someone else in the course of social interaction. Expressions of reflective self-awareness include the capacities to think and care about oneself and one's mental states or behaviour, and the most explicit forms of self-awareness entail that one has an idea or concept of self. If we discover atypicalities in self-awareness among individuals with autism, therefore, it will be important to consider whether these implicate abnormalities in individuals' sense of self, or concept of self, or both. At the same time, autism may reconfigure our view of the distinctions among different kinds of self-awareness. For example, we may discover there are forms of self/other-consciousness that antedate and underpin fully-fledged “I-thoughts” and that straddle the unreflective/reflective divide.

We begin with two classic clinical descriptions that capture how autism involves disorder in self-other relations *and* self-awareness.

Clinical descriptions

Kanner (1943) considered that the group of 11 children he described had “inborn autistic disturbances of affective contact” (p. 250). He noted how “people, so long as they left the child alone, figured in about the same manner as did the desk, the bookshelf, or the filing cabinet” (p. 246). Kanner’s case descriptions illustrate these children’s qualities of self-awareness as well as their lack of connectedness with others. For example, their seeming imperviousness to other people sometimes extended to a marked failure to respond to others calling the child’s own name, or an insensitivity to others’ attitudes to the self. Of one boy, it was reported that he “... rarely responded to any form of address, even to the calling of his name ... It made no difference whether one spoke to him in a friendly or harsh way” (p. 227–8). Kanner noted how such abnormality extended to self-expressions in language, for example, when another boy stumbled and nearly fell, and said of himself: “You did not fall down.” Kanner remarked how it was around the sixth year of life that the children he studied gradually learnt to speak of themselves in the first person and the person addressed in the second person.

Alongside these features, Kanner recorded abnormalities in non-verbal aspects of self/other awareness. For example, one child was said to move among other children “like a strange being, as one moves between the pieces of furniture in a room” (p. 241). Kanner also described a number of instances in which the children related not to what another person had just done, but to the hand that was in the way or the foot that stepped upon the child’s blocks. Yet not all aspects of the self are equally affected. When it came to the children’s attitudes toward objects, it seemed to Kanner that typically, the autistic child, “... is interested in them, can play with them happily for hours. He can be very fond of them, or get angry with them ... When with them, he has a gratifying sense of undisputed power and control” (p. 246).

Kanner also noted how the children took pleasure in achievements such as completing puzzles. Yet such pleasure was also notable for something else: of one child, Kanner wrote: “He blew out a match with an expression of satisfaction with the achievement, but did not look up to the person who had lit the match” (p. 224), and of another: “She showed no interest in test performances. The concept of test, of sharing an experience or situation, seemed foreign to her” (p. 229). It is striking how what we take to be a natural orientation toward other people’s attitudes to what we, as selves, have achieved, was notable for its absence among some of these children.

Of course, one should not take such descriptions to apply to all children with autism, especially given that the clinical picture evolves over time. Extending both the range and depth of clinical observation, Bosch (1970) illustrated how affected individuals sometimes appear to lack a sense of possessiveness as well as self-consciousness and shame, to be delayed in “acting” on others by demanding or ordering, and to be missing something of the “self-involvement, the acting with, and the identification with the acting person” (p. 81). As a reflection of this, a “delay occurs in the constituting of the other person in whose place I can put myself ... [and] ... in the constituting of a common sphere of existence, in which things do not simply refer to me but also to others” (p. 89). In these ways, Bosch framed his account of self- and other-awareness with reference to attitudes implicated in relational stances. If a child does not experience such attitudes as possessiveness or shame for him- or herself, and/or does not register and identify with the attitudes and actions of others, then that child will be deprived of a vital source of knowledge about the subjective life of other people and the shared world in which we co-exist.

A complementary perspective on the self in autism is that provided by first-hand accounts of self-experience from able adolescents and adults with autism (Frith and Happé, 1999). For instance, Grandin (1992) wrote that even in adulthood, she “had an odd lack of awareness of my oddities of speech and mannerisms until I looked at videotapes” (p. 113). Here, it is striking that only when confronted with herself depicted on videotape, rather than apprehending herself through the attitudes of others, could she become aware of her mannerisms. Moreover, as Happé (1991) has noted, Grandin’s accounts are remarkable for their lack of emphasis on her own emotional or family life, and for their portrayal of autism as an abnormality of perceptual processing and cognitive style. On the face of it, this seems to reflect not only her (probable) unengagement with others, but also her unengagement with herself-as-unengaged with others.

Having drawn on the richness of clinical descriptions, we turn to controlled studies in order to ascertain whether any abnormalities in manifestations of self-other relations are specific to individuals with autism, rather than (for example) a reflection of severe learning difficulties that also occur among children without autism. A further aim is to delineate more precisely those qualities of self-other relations and understanding that are impaired in autism, and to discover more about the kinds of self-awareness that are relatively intact.

Controlled studies of children and adolescents with autism

Relational self/other-awareness

We restrict ourselves to brief illustrations of research in the domains of person-with-person interactions and person-person-world relations. Our intention is to consider how self-other communication involves transactions between two embodied individuals who are connected with each other, yet who treat each other as separate and differentiated centres of subjectivity. This structure to interpersonal communication is critical for the development of human beings’ understanding of persons-with-minds (e.g. Hobson, 1993a,b).

We begin with an early study by Dawson, Hill, Spencer, & Galpert (1990), who studied 16 autistic children aged 2–6 years and 16 typically developing children matched for receptive language. Participants were videotaped interacting with their mothers in three different contexts: free play, a more structured situation in which the mother asked the child to help her to put away some toys, and a face-to-face situation over snack time. There were not significant group differences in the frequency or duration of gaze at the mother’s face, nor the frequency or duration of smiles in face-to-face interactions over a snack. However, children with autism were much less likely than typically developing children to combine their smiles with eye contact in a single act that seemed to convey an intent to communicate feelings. Not only this, but whereas 10 out of 14 typically developing children with codable data smiled in response to their mother’s smile, only three out of 15 children with autism ever did so. It was also observed that the mothers of the children with autism were less likely to smile in response to their children’s smiles, which after all were rarely combined with sustained eye contact.

We cite this study for the reason that it lends itself to interpretation from the point of view of self-other relations and communication. Only rarely were the children with autism seen to convey their feelings to their communicative partner, or to communicate in such a way that this appeared to be for the mother. Correspondingly, they appeared not to register or respond to their mothers’ smiles as smiles for themselves.

Or again, Wimpory, Hobson, Williams, & Nash (2000) elicited parental reports of their young children’s first 2 years of life. Children who were subsequently diagnosed as having autism contrasted with those without autism in being said to show less intense eye gaze or turn-taking with

others, and fewer expressions of greeting, anger and distress toward people. Their attitudes were not other-person-directed, nor did they take the form of self-to-other communication in the way that was reported for the young children without autism.

Additional aspects of self-other connectedness and differentiation may be illustrated by findings from a study by Sigman, Kasari, Kwon, & Yirmiya (1992). These researchers videotaped 30 young children with autism who had a mean age of under 4 years, together with matched children without autism, in the presence of an adult who appeared to hurt herself by hitting her finger with a hammer, simulated fear toward a remote-controlled robot, and pretended to be ill by lying down on a couch for a minute, feigning discomfort (and see Charman, Swettenham, Baron-Cohen, Cox, Baird, & Drew, 1997, for similar events involving 20-month-olds). In each of these situations, children with autism were unusual in rarely looking at or relating to the adult, or being affected by the adults' attitudes to the robot. These observations exemplify the children's lack of engagement with other people's attitudes toward a shared world—a world that, amongst other things, contains themselves.

It is important to emphasize that not *all* aspects of relationships are affected to the same degree among persons with autism. For instance, their attachment to a significant caregiver may be relatively intact. There are several published studies that indicate how young children with autism *do* respond to separation from and reunion with their caregivers, at least in the short-term (e.g. Rogers, Ozonoff & Maslin-Cole, 1991; Shapiro, Sherman, Calamari & Koch, 1987; Sigman & Mundy, 1989). Many (not all) 2–5-year-old children with autism are like matched learning disabled children in showing somewhat variable mood changes such as fretting when their caregiver leaves them, and upon reunion they tend to spend more time alongside the caregiver than a stranger. Therefore the children's relationship with their caregivers is clearly special, even though many qualities of their self-other relatedness are atypical (Beurkens, Hobson, & Hobson, 2013). Attachments to significant others appear to be organized by principles that differ from those that are critical for other-person-centred engagement and perspective-shifting.

Self-conscious emotions

Prototypically, self-conscious emotions reflect a person's state of self within or with respect to a social context. Feelings such as coyness, embarrassment, guilt, pride, jealousy or shame are sometimes called "social" or even "complex" emotions, on the grounds that they seem to entail sophisticated understandings of self and other people along with relatively high levels of self-consciousness. However, it is important not to pre-judge whether young children require a concept of self in order to experience this or that self-conscious emotion. At least some social emotions may have a complex structure that comes as a biological given rather than an outcome of social-cognitive development. The study of autism may help us to see the degree to which this is so.

Again we can learn much from what parents report about their children. Hobson, Chidambi, Lee, & Meyer (2006) interviewed parents of children with and without autism who were aged between approximately 6 and 13 years and matched for verbal ability (roughly that of typically developing 3–9-year-olds). Parents felt they could recognize in their children with autism not only emotions such as anger and fear, but also emotional responsiveness to other people's mood states, as well as shyness, non-person-directed pride, and jealousy. Yet seldom could they cite clear instances of other-person-centred emotions such as guilt, shame, pity, empathic concern, or embarrassment. One parent who gave a convincing account of her child's jealousy, said this about his expressions of concern: "He might be worried, but he doesn't have that empathy sort of concern—he doesn't show that at all ... Empathetic sadness isn't there."

These reports from parents dovetail with what may be gleaned from self-reports elicited from verbally fluent children and adolescents with autism. For example, Kasari, Chamberlain, & Bauminger (2001) described how high-IQ children with autism reported feeling guilt, but only 14% of participants with autism (vs. 42% of those with typical development) spoke of guilt over physical harm to others, and none referred to emotional harm such as hurting someone's feelings. When speaking of embarrassment, fewer participants with autism explicitly mentioned an audience (also Capps, Yirmiya, & Sigman, 1992). In "self-understanding interviews" conducted by Lee & Hobson (1998), children with autism were not only restricted in the feelings they expressed about themselves, but they also failed to mention friends or being members of a social group. Therefore not merely do that children and adolescents with autism seem to have difficulty in responding to another person's feelings or attitudes as belonging to that person and at times shaping their own feelings, but also this limitation is apparent in their own descriptions of what they feel.

Thirdly, there is evidence from quasi-experimental studies. Kasari, Sigman, Baumgartner, & Stipek (1993) tested matched young children with and without autism (mean age 42 months) and typically developing children of the same level of ability (mean age 23 months). Each participant completed a puzzle, and the investigator and parent reacted neutrally; then the child completed a second puzzle, and after three seconds, both adults gave praise. Although children with autism were like the comparison children in being inclined to smile when they succeeded with the puzzles, those with autism were less likely to draw attention to what they had done or to look up to an adult, and less likely to show pleasure in being praised. Their pride assumed a strangely "asocial" form. It seems that pride has two components, namely pride in accomplishing something—a feeling that is not "social"—and pride before other people.

In a similar way, Hobson et al. (2006) contrived situations in which participants might feel pride, guilt, and coyness/embarrassment. Again, children with autism were relatively less likely than matched participants without autism to manifest other-person-directed expressions of the feelings. For example, when they felt responsible for the leg falling off a doll, they were less likely to show a "guilty looks" pattern of orientation toward the tester that included expressions of relief when the tester reassured them that the doll was already broken; and when they received the attentions of a cuddly toy wielded by a playful tester, they rarely showed "re-engagement looks" that give coyness a specially intimate quality. This was despite the fact that the participants with autism showed many signs of being aware when they were the focus of attention. It seemed that there was a dissociation between these participants' self-consciousness in being observed, and their ability to be affected by and engaged with the attitudes of a particular embodied other person. This may correspond with the contrast between the ability of children with autism to remove rouge from their faces when they perceive themselves in a mirror, or indeed to recognize their bodies after a delay (Lind & Bowler, 2009), and their relative lack of coyness in that same context (Dawson & McKissick, 1984; Neuman & Hill, 1978; Reddy, Williams, Costantini, & Lan, 2010; Spiker & Ricks, 1984).

The crux is that social emotions such as concern, guilt or social pride are manifestations of affective engagement with other people as centres of subjectivity, embodied persons with whom one is linked and from whom one is differentiated (also Hobson, Harris, García-Pérez, & Hobson, 2009). Yet could not the same be said, perhaps even more strongly, of the emotion of jealousy? There is good evidence, both from parent report in the research by Hobson et al. (2006) already cited, and also from experimental studies conducted by Bauminger (2004; Bauminger, Chomsky-Smolkin, Orbach-Caspi, Zachor, & Levy-Shiff, et al., 2007), that children with autism manifest and experience jealousy. Surely this contradicts the idea that their self-other relatedness is impaired?

Not necessarily. There may be more diversity among social emotions than we have realized. Perhaps jealousy is a biologically prepared emotional state tied in with processes of attachment rather than intersubjectivity, and like attachment, relatively spared among individuals with autism. This is not to suggest that attachment relationships are irrelevant for the development of self-experience. Indeed, self-related aspects of feelings of jealousy might make a contribution to the self-awareness of children with autism. Dissociations among different social emotions in autism highlight the possibility that certain seemingly complex emotions (including but not necessarily restricted to jealousy) do not require as much as we might have imagined by way of intersubjective experience or cognitive-conceptual sophistication.

Imitation

The reason that so much attention has been paid to young children's developing propensities and abilities to imitate other people, is that here we may find not only reflections of developing self-other awareness, but also pointers to the mechanisms through which new levels of self-other understanding might be achieved. There is a complex and in part conflicting literature on this topic. On the one hand, there are many clinical and experimental reports to indicate that children with autism find it hard and/or are rarely moved to imitate a range of emotional expressions, bodily movements, and pantomimed actions of other people (e.g. DeMyer, Alpern, Barton, DeMyer, Churchill, Hingtgen, et al., 1972; Rogers, Hepburn, Stackhouse, & Wehner, 2003). On the other hand, children with autism are able to copy the goal-directed actions of someone else (e.g. Charman & Baron-Cohen, 1994), and are prone to "echo" the behavior of others. Moreover, several studies have reported how children with autism show responsiveness to being imitated, so that they often become more socially engaged and interactive when an adult imitates their actions (Dawson & Adams, 1984; Dawson & Galpert, 1990). Therefore the specific qualities of the children's imitative deficits may betray something about the basis for their limitations in self-other awareness, and the specific qualities of their imitative abilities illuminate how they develop some forms of self-consciousness and self-concept.

Two aspects of imitation are especially revealing, insofar as they appear to tap the process of identifying with someone else. Hobson & Lee (1999) tested matched groups of children with and without autism aged between 9 and 19 years (and verbal mental ages between 4 and 13 years) for their ability to imitate a person demonstrating four novel goal-directed actions on objects in two contrasting "styles." In one condition, the demonstrator made a toy policeman-on-wheels move along by pressing down on its head either with his wrist cocked or with extended index and middle fingers. In other conditions, he showed either gentle or harsh styles of action. The results were that children with autism were perfectly able to copy the demonstrator's actions, for example in pressing down the policeman's head to make him move, but contrasted with control participants insofar as very few adopted the demonstrator's style of acting upon the objects. Instead of adopting the wrist or two-finger approach to activating the toy, for example, most of them pressed down on the policeman's head with the palm of a hand. Here there appeared to be a contrast between children's ability to observe and copy intended actions *per se*, relatively intact in autism, and the propensity to identify with and thereby imitate a **person's** expressive mode of relating to objects and events in the world. Secondly, when the demonstrator held a pipe-rack against his own shoulder in order to strum it with a stick, a substantial majority of the control participants positioned the pipe-rack against their own shoulder before strumming it, but most of the children with autism positioned the pipe-rack at a distance in front of them, on the table. Again with respect to self-orientation, the children with autism did not identify with the other and perform the actions from a person-centered perspective (also Meyer & Hobson, 2004).

Awareness of oneself as having mental states

The most systematic body of controlled experimental research on self-awareness of mental states comes from work in the “theory-of-mind” tradition. Early empirical research in this area included that by Perner, Frith, Leslie, & Leekam (1989), who reported that children with autism found difficulty in judging whether or not they themselves (and not only an experimenter) knew what was inside a container, on the basis of whether they had had an opportunity to look inside the container for themselves. In the late 1990’s, Frith & Happé (1999) offered an overview and theoretical interpretation of such research, augmenting experimental evidence with examples of impoverished introspective self-reports from three men with Asperger syndrome (originally Hurlburt, Happé, & Frith, 1994) as well as first-person accounts by writers who have autism. For these authors, self-consciousness “may be seen as the product of a specific neurocognitive mechanism” (p. 18) that they take to underlie theory-of-mind abilities. However, it is plausible that causative arrows fly in the opposite direction: a young child’s experiences of engagement with others’ attitudes toward the self and a shared world may constitute the psychological “mechanism” for developing self-reflective awareness and acquiring mental state concepts.

A careful review by Williams (2010) brings us up-to-date on the sources of evidence that indeed, individuals with autism are impaired in recognizing their own mental states. Two examples from Williams’ own work in collaboration with Happé serve to illustrate recent studies. First, Williams & Happé (2009a) returned to re-examine a result from the earlier study by Perner et al. (1989), suggesting that children with autism found it easier to report their own previously held false belief about the contents of a Smarties tube than to identify the false belief of someone else. Williams & Happé (2009a) noted that here, participants had been asked to make explicit statements about their belief over the contents of the tube—a belief that subsequently turned out to be false—prior to being asked what that early belief had been. Therefore they might simply have responded by drawing upon what they recalled they had said, rather than needing to recall what they had believed. Williams & Happé (2009a) introduced a novel design in which a tester pretended to have cut a finger and asked participants to fetch a plaster (Band-Aid). Participants had to choose from among three boxes, only one of which was a plasters box (but, in fact, contained candles), thereby demonstrating but not stating their false belief that this would contain plasters. When subsequently asked what they had thought was in this box, participants with autism showed more difficulty in answering this question than a similar question framed in relation to what another person would think. The authors concluded that children with autism may be relatively more impaired at recognizing their own vis-à-vis others’ mental states of belief.

A second study by Williams & Happé (2010) was designed to contribute to somewhat conflicting evidence (Phillips, Baron-Cohen, & Rutter, 1998; Russell & Hill, 2001) on whether children with autism can recognize their own intentions. Participants were given a “knee-jerk task” (Lang & Perner, 2002) in which they were asked if they had intended to move their leg when, in fact, a knee reflex had been elicited by the tester. The results were that compared with matched participants without autism, those with autism more often claimed that their reflex movement had been under their intentional control. Such diminished awareness of their own intentions was related to their ability to recognize others’ mental states, as assessed by performance on false belief tasks.

If awareness of oneself as having mental states is compromised in autism, then what influence might this have on remembering, and perhaps especially recalling personally experienced events? Lind (2010) has provided an excellent review of research on various aspects of self-experience among individuals with autism, and in particular, the relation between atypicalities in self-experience and memory. Lind makes the point that having a concept of self might be important for encoding

and retrieving personally significant memories, yet it is also the case that the ability to have lively memories of one's past personal experiences makes a vital contribution to one's notion and sense of self. She discusses how partial self-knowledge and self-concepts among children with autism are reflected in impairments in autobiographical episodic memory. The evidence includes impoverished accounts of specific personal experiences in participants' reports of task-related events, natural everyday happenings, and past aspects of their lives (e.g. Bruck London, Landa, & Goodman, 2007; Crane & Goddard, 2008; Goddard, Howlin, Dritschel, & Patel, 2007; Klein, Chan, & Loftus, 1999; Losh & Capps, 2003; Millward, Powell, Messer, & Jordan, 2000). Their restricted non-autobiographical episodic memory, for instance in "remembering" things rather than simply "knowing" them (Bowler, Gardiner, & Grice, 2000), may also attest to deficits in re-experiencing themselves as a subject of experience. Indeed, these limitations may extend to imagining a future self, and amount to a restriction on what Lind calls the "temporally extended self-concept" (Lind, 2010, p. 430).

In addition to this, there is tentative but growing evidence that children with autism may not show the usual enhanced processing of information that is encoded in relation to the self (e.g. Henderson, Zahka, Kojkowski, Inge, Schwartz, Hileman, et al., 2009; Lombardo, Barnes, Wheelwright, & Baron-Cohen, 2007; Toichi, Kamio, Okada, Sakihama, Youngstrom, Findling, et al., 2002). Or again, to judge from their drawings, they may experience themselves as relatively undifferentiated human figures (Lee & Hobson, 2006). Yet it is important to remember that some features of self-experience appear to be relatively intact among individuals with autism, even if these often although not exclusively focus on semantic (rather than episodic), physical (rather than psychological), and individual (rather than socially embedded) characteristics (e.g. Lee & Hobson, 1998).

Communication and language

Here, we come full-circle, to consider self-other relations in verbal as well as non-verbal communication.

If a child is to adjust his or her language according to situational and communicative context as construed by conversational partners, then that individual needs to co-ordinate linguistic expressions with what he or she interprets to be the perspectives expressed and anticipated by the partner. More than this, he or she needs the propensity to engage with the other's perspective in such a way as to make the appropriate adjustments, a motivational as well as a cognitive matter. Therefore, if children with autism are relatively unengaged with other people's attitudes, it might be expected that they would show limited in sensitivity to pragmatic adjustments in language.

Consider first a study of one aspect of non-verbal communication. Hobson & Meyer (2005) devised an original methodology (the "Sticker Test") and demonstrated that whereas children without autism would often employ a point-to-themselves (i.e. a location on their own body) to communicate that a tester should place a sticker on herself (i.e. the corresponding location on her body), this was much less frequently the case among children with autism. Here the children without autism appeared to identify with the tester in assuming that she would interpret the child's self-orientated action as one with which she should identify, in order to place the sticker on her own (i.e. the tester's) body. Participants with autism seldom adjusted their communication in this mutually coordinated, person-anchored way.

An intimately related aspect of the pragmatics of verbal language is the comprehension and use of deictic terms. Deictic terms such as "here" and "there" or "this" and "that"—as well as "I" and "you"—have meanings that are anchored in the embodied stances of speaker and listener. From his observations of children with autism, Kanner concluded that personal pronouns "*are repeated*

just as heard, with no change to suit the altered situation” (Kanner, 1943, p. 244). Instead of relating the other person’s utterance to that person’s attitude and then identifying with the other person’s stance, children with autism tend to adopt speech forms that correspond with *their* experience of the circumstances in which the words are uttered, and therefore to repeat utterances as heard (Charney, 1981; Jordan, 1989; Lee, Hobson, & Chiat, 1994). This represents a failure to recognize and assume the other person’s attitude-in-speaking. Indeed, in a study by Loveland & Landry (1986), correct production of I/you pronouns by autistic children was related to the number of their spontaneous initiations of joint attention with an experimenter. This suggests that correct usage of deictic terms and pronouns may reflect a special quality of engagement and co-reference between self and other.

Evidence compatible with this account of atypical self- and other-reference comes from studies by Jordan (1989) and Lee et al. (1994), where in settings such as being the object of a puppet’s tickling, or when referring to photographs of themselves, children with autism would sometimes give proper names to themselves or the experimenter sitting alongside, rather than using the pronouns “me” or “you.” Participants with autism had a relatively detached, almost third-person attitude to photographs of themselves and the experimenter. In contrast, children without autism seemed to identify with the depictions of themselves, and to see and care about the photographed person as “me.”

In a study of deictic communication (Hobson, García-Pérez, & Lee, 2010), we employed semi-structured tests to determine whether children with autism produce and comprehend such person-centred expressions. In several respects, they behaved rather like matched comparison children without autism, but there were also subtle and telling group differences. In particular, a majority of children with autism, but not a single child in the comparison group, sometimes referred to a location that was distant from themselves with the terms “this” or “here” (rather than “that” or “there”), or pointed with unusual precision with what we came to call a “laser-beam point” that was sometimes accompanied by lining up an eye behind the look. Not only this, but also participants with autism were less likely to accompany points with a look back to the person for whom the points should have been intended.

These findings show us something further about atypicalities in self-other relations among children with autism. In the typical case, a point is understood by a listener with reference to current discourse, so that it is not necessary for a speaker to be exact in conveying what is meant, only precise enough to communicate which of several alternatives is the referent singled out. “This” and “here” are terms used in relation to speaker-listener locations and the topic of discourse, so that “this” can refer to an immediately proximal location, a room, a town, a country, and so on. Yet when the children with autism made overly specific “laser-beam points,” or when they used the word “this” to refer to a distal location, they appeared to be operating within an egocentric framework rather than one that had reference to common ground (Clark, 1996) shared between themselves and the tester. Whereas participants without autism mostly looked back to the person for whom their points were intended and framed, such looks were less frequent among those with autism. An additional finding was that some children with autism found it difficult to appreciate the meaning of the tester’s atypical gesture (a head-nod to indicate location) when this was intended as a communication for the children themselves. Each atypicality appears to reflect children’s limited co-ordination of interpersonal experience and reciprocal role-taking in relation to a shared world experienced jointly in common with others.

Neuroscientific perspectives

One neurofunctional approach attempting to capture atypicalities in preconceptual self-other relations is that concerned with the operation of “mirror neurones” (e.g. Decety & Chaminade,

2003; Gallese, 2001). Although there is some evidence, both from fMRI findings (Dapretto, Davies, Pfeifer, Scott, Sigman, & Bookheimer, et al., 2006) and EEG patterns of mu frequency suppression (Oberman, Hubbard, McCleery, Altschuler, Ramachandran, & Pineda, 2005) that such functioning may be atypical among children with autism, the claims of the “mirror neuron” theory of autism (Williams, Whiten, Suddendorf, & Perrett, 2001) are disputed on both empirical and theoretical grounds (e.g. Southgate & Hamilton, 2008).

Another important perspective on self-other connectedness and differentiation is that elaborated by Lombardo and Baron-Cohen (2011), who have combined a theoretical elaboration of ideas from social psychology with neurofunctional investigations. These authors emphasize the importance of comparing how individuals with autism thinking about themselves on the one hand (self-referential processing), and thinking about other people on the other. In the former respect, they stress how little is known about impairments in sources of information about the self among persons with autism, for instance through interoceptive (e.g. somatosensory or visceral) information (also Mundy, Sullivan, & Mastergeorge, 2009; Silani, Bird, Brindley, Singer, Frith, & Frith, 2008). They illustrate the importance of this issue through a transcranial magnetic stimulation study in which they reported that among participants with Asperger syndrome, corticospinal excitability to viewing another’s pain was reduced (Minio-Paluello, Baron-Cohen, Avenanti, Walsh, & Aglioti, 2009). They consider the possibility that such individuals may be aroused by such appraisals (Ben Shalom, Mostofsky, Hazlett, Goldberg, Landa, Faran, et al., 2006), but unable to draw on such information to feel empathy.

There is also a body of work concerned with the functioning of those parts of the brain that appear to be implicated in thinking about self and other. For example, Lombardo et al. (2007) have reported that the right temporo-parietal junction, which is thought to be a brain region subserving the representation of mental states, was hypoactive among participants with ASD when mentalizing about self and others. On the other hand, in a study of brain functioning while participants reflected on themselves or others (Lombardo et al., 2009), the ventromedial prefrontal cortex responded atypically, in what the authors describe as “an egocentrically equivalent fashion for both self and other” (Lombardo & Baron-Cohen, 2011, p. 137). Moreover, the degree to which this part of the prefrontal cortex responded most “egocentrically” to the mental characteristics of self corresponded with participants’ degree of social impairment in early childhood.

Lombardo and Baron-Cohen have argued that if self-referential processing is important for explaining some of the mechanisms involved in mindblindness among individuals with autism, one would expect there to be specific areas of the brain that respond atypically for thinking about the mental states specifically of oneself. When typical participants are asked to “mentalize” about themselves and others, for instance, they show increased activation in the middle cingulate cortex when the focus is themselves, rather than others; when participants with autism were asked to do this, they responded more when thinking about the mental states of others vis-à-vis themselves (Lombardo, Chakrabarti, Bullmore, Sadek, Pasco, Wheelwright, et al., 2010). One conclusion drawn by the authors is that there cannot be one general-purpose mechanism accounting for all deficits in mentalizing about oneself and others.

In view of the attention that such neurofunctional studies have attracted, it is important to remember that they need to be interpreted with caution. We simply do not know whether atypicalities in neurological functioning among individuals beyond early childhood underlie—and point to the mechanisms of—psychological abnormality, or whether instead they reflect the outcome of developmental processes that might originate elsewhere. There are complications over the relations between the development of understandings of mind (or “mentalizing”) on the one hand, and the mechanisms of self-other connectedness and differentiation on the other. In particular,

basic processes of intersubjective engagement may provide necessary foundations for mental state concepts, and thinking involving these concepts might become a feature of partly localized neurological functioning.

Autism and the development of self

One lesson to be drawn from the clinical and experimental evidence we have cited, is that there are several strands to the story of the development of self in typical and atypical development. We have seen there is substantial evidence that individuals with autism have limitations in self-awareness. The evidence comes from domains as diverse as self-other relations and social emotions, imitation, “theory-of-mind,” memory, and linguistic functioning, and is complemented by evidence from neurofunctional studies (also Hobson, 2010; Lombardo & Baron-Cohen, 2010). Although much (not all) of the most decisive evidence pertains to features of self-awareness that bear an intimate connection to the children’s social relations, it remains unclear how far the children’s limitations in self-awareness reflect their atypical engagement with and restricted understanding of other people.

In particular, we have given little consideration to possible abnormalities in the children’s pre-reflective sense of self, in relation to their own bodily experience. On the one hand, there are grounds for believing that certain forms of bodily self-experience are intact. For instance, Williams & Happé (2009b) reported that children with autism were proficient in a test in which they positioned a mouse to move one of several coloured squares on a computer screen, and had to judge which was under their intentional control and which “distractor” squares were under the control of the computer, suggesting no impairment in their sense of physical agency. On the other hand, there is a need to account for individuals’ reports of bodily hypo- or hyper-sensitivity to sound, light, touch or pain (Minshew & Hobson, 2008). Are these simply a reflection of an inability to reflect on inner experiential states (Frith & Happé, 1999), or do they betray something more basic about self-experience and self-other relations? There are related, seemingly “basic” abnormalities in the social sphere. Hobson et al. (2006) report that in every one of a (modestly sized) group of children with autism, parents described abnormalities in their offspring’s sense of bodily space. Here is one illustrative vignette:

“He does that, he bumps into people in the supermarket. He’s so unaware of other people, if someone’s looking at something on a shelf he’ll go in between them and the shelf ... it’s just like everybody’s an object”

Drawing on phenomenological considerations and self-reports from people with autism, Farley (2010) has argued that individuals with autism have profoundly unusual experience of their own bodies, and that this renders them both prone to disconnection from, and over-sensitive to, their social and non-social environment. From a neuroscientific perspective, Mundy, Gwaltney, & Henderson (2010) suggest that autism involves early impairments in the capacity for rapid integrated processing of proprioceptive and interoceptive (self-referenced) information on the one hand, and other-referenced (exteroceptive) information on the other, as manifest, for example, in failures of joint attention. It may prove that the distinctions we make between individual and social experience, between physical and psychological contents to self-awareness, and between non-reflective and reflective self-awareness, are far from straightforward.

Having said this, it is worth returning to what we have discovered about the atypicalities of self-other experience among individuals with autism in the social domain. In summary, children with autism have a relative dearth of engagement with other people’s feelings *as* located in the other people and of importance for themselves in one way or another. This importance might

either take the form of concern for the other, or for themselves in the eyes of the other, or for what the world means for the other and therefore what it might mean for the self. It is not that they fail to react to others' expressions of feeling, for instance, but often they seem to lack the self-other organization of attitude and action. Neither are they gripped by other people, nor are they so easily moved to assimilate or adjust to the stance of someone else, whether in settings of social referencing or imitation or communication. Along with this, as Bosch (1970) suggested long ago, they find difficulty in constituting a "common sphere of existence" with other people. Frith & De Vignemont (2005) offer an interesting elaboration of such a view, suggesting that people with Asperger syndrome suffer from a disconnection between a strong naïve egocentric stance, where the other person is represented in relation to the self, and an allocentric stance detached from interactions with people, where the existence or mental states of other people need to be represented as independent from the self.

The question that all this raises is whether we need to introduce structures of self/other connectedness and differentiation into our account of the most basic forms of human social experience, and whether the processes that so organize social relations are weak or missing among children who develop autism. We propose that the answer to this question is "Yes." (This is an over-simplification, because we believe that the developmental basis for autism is a systemic breakdown in person-person-world relations that can be affected by other factors such as congenital blindness, but these are exceptional cases). More specifically, and from early in life, typically developing individuals identify with other people (Freud, 1955/21). In identifying with someone else, the self not only responds to another individual's bodily-expressed orientation from that other person's stance, but also assimilates that orientation so that it becomes a possible mode of relating for the self.

In order to understand some of the profound implications of this process for typical development, consider how one-year-olds' acts of showing or pointing out things to other people reflects their engagement with others as separate sources of attitude to objects and events in the world. Moreover, identifying with the attitudes of others is a primary way to establish a connection between first-person phenomenological experience of, say, feelings of distress or possessiveness or agency, and other people's experiences of these kinds (Barresi & Moore, 1996). This kind of non-inferential and pre-conceptual process lays the foundations for developments around the middle of the second year of life, when children acquire the abilities to conceptualize what it means for people to have their own "selves" and psychological stances, to exercise self-reflective awareness, and to introduce (originally person-anchored) perspectives to new objects in symbolic play (Hobson, 1993, 2002).

If it is the case that this form of biologically given structuring of social engagement is not fully in place for some children, their acquisition of concepts of self and others (with minds) will be compromised to the degree that they miss out on preconceptual forms of experience of relations with embodied, and bodily expressive, other people. If they are seldom affected (through identification) by the other person's attitudes as the attitudes of another self with whom they are engaged—or indeed, if they are limited in identifying with themselves as experienced in the past and projected into the future (also Farley, 2010)—they will be seriously constrained in developing the full range of social emotions, in imitating others, in developing autobiographical and other forms of episodic memory, in making pragmatic adjustments in communication, and in acquiring many other aspects of self-awareness. Yet as we have seen, many children with autism have some concept of self and some capacity to acquire self-reflection (although there are no studies on whether this ability is hard-won and achieved relatively late in childhood). Their concepts of self and their range of self-directed attitudes are limited in virtue of their abnormality in some, but only some, of several dissociable lines of development that contribute to the typical development of self.

Acknowledgement

Some sections of this paper are adapted from a chapter entitled “Autism: Self and other,” to appear in S. Gallagher (Ed.), *Oxford Handbook of the Self* (2012), by permission of Oxford University Press, Inc. We gratefully acknowledge the support of the Baily Thomas Charitable Trust. This chapter was written while the first author was a Fellow and the second author was a Visiting Scholar at the Center for Advanced Study in Behavioural Sciences (CASBS) at Stanford University.

References

- Barresi, J., & Moore, C. (1996). Intentional relations and social understanding. *Behavioral and Brain Sciences* 19: 107–54.
- Bauminger, N. (2004). The expression and understanding of jealousy in children with autism. *Development and Psychopathology* 16: 157–77.
- Bauminger, N., Chomsky-Smolkin, L., Orbach-Caspi, E., Zachor, D., & Levy-Shiff, R. (2007). Jealousy and emotional responsiveness in young children with ASD. *Cognition and Emotion* 22: 595–619.
- Ben Shalom, D., Mostofsky, S. H., Hazlett, R. L., Goldberg, M. C., Landa, R. J., Faran, Y. et al. (2006). Normal physiological emotions, but differences in expression of conscious feelings in children with high-functioning autism. *Journal of Autism and Developmental Disorders* 36: 395–400.
- Beurkens NM, Hobson JA, Hobson RP. (2013). *Autism severity and qualities of parent-child relations*. *Journal of Autism and Developmental Disorders*, 43,168–178.
- Bosch, G. (1970). *Infantile Autism* (transl. D. Jordan & I. Jordan). New York: Springer-Verlag.
- Bowler, D. M., Gardiner, J. M., & Grice, S. (2000). Episodic memory and remembering in adults with Asperger's syndrome. *Journal of Autism and Developmental Disorders* 30: 305–16.
- Bruck, M., London, K., Landa, R., & Goodman, J. (2007). Autobiographical memory and suggestibility in children with autism spectrum disorder. *Development and Psychopathology* 19: 73–95.
- Capps, L., Yirmiya, N., & Sigman, M. (1992). Understanding of simple and complex emotions in non-retarded children with autism. *Journal of Child Psychology and Psychiatry* 33: 1169–82.
- Charman, T., & Baron-Cohen, S. (1994). Another look at imitation in autism. *Development and Psychopathology* 6: 403–13.
- Charman, T., Swettenham, J., Baron-Cohen, S., Cox, A., Baird, G., & Drew, A. (1997). Infants with autism: An investigation of empathy, pretend play, joint attention, and imitation. *Developmental Psychology* 33: 781–9.
- Charney, R. (1981). Pronoun errors in autistic children: Support for a social explanation. *British Journal of Disorders of Communication* 15: 39–43.
- Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.
- Crane, L., & Goddard, L. (2008) Episodic and semantic autobiographical memory in adults with autism spectrum disorders. *Journal of Autism and Developmental Disorders* 38: 498–506.
- Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., et al. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* 9: 28–30.
- Dawson G., & Adams, A. (1984). Imitation and social responsiveness in autistic children. *Journal of Abnormal Child Psychology* 12: 209–26.
- Dawson, G., & Galpert, L. (1990). Mother's use of imitative play for facilitating the social behavior of autistic children. *Development and Psychopathology* 2: 151–62.
- Dawson, G., Hill, D., Spencer, A., Galpert, L., & Watson, L. (1990). Affective exchanges between young autistic children and their mothers. *Journal of Abnormal Child Psychology* 18: 335–45.
- Dawson, G., & McKissick, F. C. (1984). Self-recognition in autistic children. *Journal of Autism and Developmental Disorders* 14: 383–94.

- Decety, J., & Chaminade, T. (2003). When self represents the other: A new cognitive neuroscience view on psychological identification. *Consciousness and Cognition* 12: 577–96.
- DeMyer, M. K., Alpern, G. D., Barton, S., DeMyer, W. E., Churchill, D. W., Hingtgen, J. N., Bryson, C. Q., Pontius, W., & Kimberlin, C. (1972.) Imitation in autistic, early schizophrenic, and non-psychotic subnormal children. *Journal of Autism and Childhood Schizophrenia* 2: 264–87.
- Farley, A. (2010). *Bodily Experience, Intersubjectivity and Re-experiencing in Autism*. Poster presented at conference on Embodiment, intersubjectivity, and psychopathology, Heidelberg, Germany, September.
- Freud, S. (1955/1921). Identification. In: J. Strachey (Ed.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. xviii, pp. 105–10. London: Hogarth.
- Frith, U., & Happé, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic? *Mind and Language* 14: 1–22.
- Frith, U., & de Vignemont, F. (2005). Egocentrism, allocentrism, and Asperger syndrome. *Consciousness and Cognition* 14: 719–38.
- Gallese, V. (2001). The “shared manifold” hypothesis: From mirror neurons to empathy. *Journal of Consciousness Studies* 8: 33–50.
- Goddard, L., Howlin, P., Dritschel, B., & Patel, T. (2007) Autobiographical memory and social problem-solving in Asperger syndrome. *Journal of Autism and Developmental Disorders* 37: 291–300.
- Grandin, T. (1992). An inside view of autism. In E. Schopler & G. B. Mesibov (Eds), *High-functioning Individuals with Autism* (pp 105–26). New York: Plenum Press.
- Happé, F. G. E. (1991). The autobiographical writings of three Asperger syndrome adults: Problems of interpretation and implications for theory. In U. Frith (Ed.), *Autism and Asperger Syndrome* (pp. 207–42). Cambridge: Cambridge University Press.
- Henderson, H. A., Zahka, N. E., Kojkowski, N. M., Inge, A. P., Schwartz, C. B., Hileman, C. M., et al. (2009) Self-referenced memory, social cognition, and symptom presentation in autism. *Journal of Child Psychology and Psychiatry* 50: 853–61.
- Hobson, J. A., Harris, R., García-Pérez, R., & Hobson, R. P. (2009). Anticipatory concern: A study in autism. *Developmental Science* 12: 249–63.
- Hobson, R. P. (1990). On the origins of self and the case of autism. *Development and Psychopathology* 2: 163–81.
- Hobson, R. P. (1993a). *Autism and the Development of Mind*. Hove: Erlbaum.
- Hobson, R. P. (1993b). The emotional origins of social understanding. *Philosophical Psychology* 6: 227–49.
- Hobson, R. P. (2002). *The Cradle of Thought*. London: Macmillan/New York, Oxford University Press.
- Hobson, R. P. (2010). Explaining autism: Ten reasons to focus on the developing self. *Autism* 14: 391–407.
- Hobson, R. P., Chidambi, G., Lee, A., & Meyer, J. (2006). Foundations for self-awareness: An exploration through autism. *Monographs of the Society for Research in Child Development* 284: 1–165.
- Hobson, R. P., García-Pérez, R., & Lee, A. (2010). Person-centred (deictic) expressions and autism. *Journal of Autism and Developmental Disorders* 40: 403–15.
- Hobson, R. P., & Lee, A. (1999). Imitation and identification in autism. *Journal of Child Psychology and Psychiatry* 40: 649–59.
- Hobson, R. P., & Meyer, J. A. (2005). Foundations for self and other: A study in autism. *Developmental Science* 8: 481–91.
- Hurlburt, R. T., Happé, F., & Frith, U. (1994). Sampling the form of inner experience in three adults with Asperger syndrome. *Psychological Medicine* 24: 385–95.
- Jordan, R. R. (1989). An experimental comparison of the understanding and use of speaker-addressee personal pronouns in autistic children. *British Journal of Disorders of Communication* 24: 169–79.
- Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child* 2: 217–50.
- Kasari, C., Chamberlain, B., & Bauminger, N. (2001). Social emotions and social relationships: Can children with autism compensate? In J. A. Burack, T. Charman, N. Yirmiya, & P. R. Zelazo (Eds), *The Development of Autism* (pp. 309–23). Mahwah: Erlbaum.

- Kasari, C., Sigman, M. D., Baumgartner, P., & Stipek, D. J. (1993). Pride and mastery in children with autism. *Journal of Child Psychology and Psychiatry* 34: 352–62.
- Klein, S. B., Chan, R. L., & Loftus, J. (1999). Independence of episodic and semantic self-knowledge: The case from autism. *Social Cognition* 17: 413–36.
- Lang, B., & Perner, J. (2002). Understanding of intention and false belief and the development of self-control. *British Journal of Developmental Psychology* 20: 67–76.
- Lee, A., & Hobson, R. P. (1998). On developing self-concepts: A controlled study of children and adolescents with autism. *Journal of Child Psychology and Psychiatry* 39: 1131–41.
- Lee, A., & Hobson, R. P. (2006). Drawing self and others: How do children with autism differ from those with learning difficulties? *British Journal of Developmental Psychology* 24: 547–65.
- Lee A., Hobson, R. P., & Chiat, S. (1994). I, you, me and autism: An experimental study. *Journal of Autism and Developmental Disorders* 24: 155–76.
- Lind, S. E. (2010). Memory and the self in autism. *Autism* 14: 430–56.
- Lind, S., & Bowler, D. M. (2009). Delayed self-recognition in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders* 39: 643–50.
- Lombardo, M. V., Barnes, J. L., Wheelwright, S. J., & Baron-Cohen, S. (2007). Self-referential cognition and empathy in autism. *PLoS One* 2: e883.
- Lombardo, M. V., & Baron-Cohen, S. (2010). Unraveling the paradox of the autistic self. *WIREs Cognitive Science* 1: 393–403.
- Lombardo, M. V., & Baron-Cohen, S. (2011). The role of the self in mindblindness in autism. *Consciousness and Cognition* 20: 130–40.
- Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Sadek, S. A., Pasco, G., Wheelwright, S. J., Suckling, J., MRC AIMS Consortium, & Baron-Cohen, S. (2010). Atypical neural self-representation in autism. *Brain* 133: 611–24.
- Losh, M., & Capps, L. (2003) Narrative ability in high-functioning children with autism or Asperger's syndrome. *Journal of Autism and Developmental Disorders* 33: 239–51.
- Loveland, K. A., & Landry, S. H. (1986). Joint attention and language in autism and developmental language delay. *Journal of Autism and Developmental Disorders* 16: 335–49.
- Meyer, J. A., & Hobson, R. P. (2004). Orientation in relation to self and other: The case of autism. *Interaction Studies* 5: 221–44.
- Millward, C., Powell, S., Messer, D., & Jordan, R. (2000). Recall for self and other in autism: Children's memory for events experienced by themselves and their peers. *Journal of Autism and Developmental Disorders* 30: 15–28.
- Minio-Paluello, I., Baron-Cohen, S., Avenanti, A., Walsh, V., & Aglioti, S. M. (2009). Absence of embodied empathy during pain observation in Asperger syndrome. *Biological Psychiatry* 65: 55–62.
- Minshew, N.J., & Hobson, J.A. (2008). Sensory sensitivities and performance on sensory perceptual tasks in high-functioning individuals with autism. *Journal of Autism and Developmental Disorders* 38: 1485–1498.
- Mundy, P., Gwaltney, M., & Henderson, H. (2010). Self-referenced processing, neurodevelopment and joint attention in autism. *Autism* 14: 408–29.
- Mundy, P., Sullivan, L., & Mastergeorge, A. M. (2009). A parallel and distributed-processing model of joint attention, social cognition and autism. *Autism Research* 2: 2–21.
- Neuman, C. J., & Hill, S. D. (1978). Self-recognition and stimulus preference in autistic children. *Developmental Psychobiology* 11: 571–8.
- Oberman, L. M., Hubbard, E. M., McCleery, J. P., Altschuler, E. L., Ramachandran, V., & Pineda, J. A. (2005). EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Cognitive Brain Research* 24: 190–8.
- Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Explorations of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development* 60: 689–700.

- Phillips, W., Baron-Cohen, S., & Rutter, M. (1998). Understanding intention in normal development and autism. *British Journal of Developmental Psychology* 16: 337–48.
- Reddy, V., Williams, E., Costantini, C., & Lan, B. (2010). Engaging with the self: Mirror behavior in autism, Down syndrome and typical development. *Autism* 14: 531–46.
- Rogers, S. J., Hepburn, S. L., Stackhouse, T., & Wehner, E. (2003). Imitation performance in toddlers with autism and those with other developmental disorders. *Journal of Child Psychology and Psychiatry* 44: 763–81.
- Rogers, S. J., Ozonoff, S., & Maslin-Cole, C. (1991). A comparative study of attachment behavior in young children with autism or other psychiatric disorders. *Journal of the American Academy of Child and Adolescent Psychiatry* 30: 483–8.
- Russell, J., & Hill, E. L. (2001). Action monitoring and intention reporting in children with autism. *Journal of Child Psychology and Psychiatry* 42: 317–28.
- Shapiro, T., Sherman, M., Calamari, G., & Koch, D. (1987). Attachment in autism and other developmental disorders. *Journal of the American Academy of Child and Adolescent Psychiatry* 26: 485–90.
- Sigman M., & Mundy, P. (1989). Social attachments in autistic children. *Journal of the American Academy of Child and Adolescent Psychiatry* 28: 74–81.
- Sigman, M. D., Kasari, C., Kwon, J. H., & Yirmiya, N. (1992). Responses to the negative emotions of others by autistic, mentally retarded, and normal children. *Child Development* 63: 796–807.
- Silani, G., Bird, G., Brindley, R., Singer, T., Frith, C., & Frith, U. (2008). Levels of emotional awareness and autism: An fMRI study. *Social Neuroscience* 3: 97–112.
- Southgate, V., & Hamilton, A. F. de C. (2008). Unbroken mirrors: challenging a theory of autism. *Trends in Cognitive Sciences* 12: 225–9.
- Spiker, D., & Ricks, M. (1984). Visual self-recognition in autistic children: Developmental relationships. *Child Development* 55: 214–25.
- Toichi, M., Kamio, Y., Okada, T., Sakihama, M., Youngstrom, E. A., Findling, R. L., et al. (2002) A lack of self-consciousness in autism. *American Journal of Psychiatry* 159: 1422–4.
- Williams, D. (2010). Theory of own mind in autism. *Autism* 14: 474–94.
- Williams, D. M., & Happé, F. (2009a). What did I say? vs. What did I think? Attributing false beliefs to self amongst children with and without autism. *Journal of Autism and Developmental Disorders* 39: 865–73.
- Williams, D. M., & Happé, F. (2009b). Pre-conceptual aspects of self-awareness in Autism Spectrum Disorder: The case of action-monitoring. *Journal of Autism and Developmental Disorders* 39: 251–9.
- Williams, D. M., & Happé, F. (2010). Representing intentions in self and others: studies of autism and typical development. *Developmental Science* 13: 307–19.
- Williams, J. H. G., Whiten, A., Suddendorf, T., & Perrett, D. I. (2001). Imitation, mirror neurons and autism. *Neuroscience & Biobehavioral Reviews* 25: 287–95.
- Wimpory, D. C., Hobson, R. P., Williams, J. M., & Nash, S. (2000). Are infants with autism socially engaged? A study of recent retrospective parental reports. *Journal of Autism and Developmental Disorders* 30: 525–36.

A review of theory of mind interventions for children and adolescents with autism spectrum conditions

Julie A. Hadwin and Hanna Kovshoff

Overview

Theory of mind (ToM) reflects an understanding that people have mental states (desires, beliefs, intentions) that are linked to feelings and behavior (Baron-Cohen, 2000). Typically, children 3–4 years of age start to show some ability to recognize that they themselves and others have beliefs, that these are sometimes false and that a person will act on them irrespective of the reality of a situation (Wimmer & Perner, 1983). The measurement of false belief (or first-order ToM) is argued to be a key marker in children's understanding of mental states (Perner, 1993) and its emergence at this age in development has been characterized as a conceptual shift in children's thinking (Wellman, Cross, & Watson, 2001). Further research has found that the development of ToM can be advanced to some extent by a favorable social environment (McElwain & Volling, 2004; Perner, Ruffman, & Leekam, 1994) or more advanced language skills (Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991). In addition, researchers increasingly recognize that some aspects of ToM (e.g. joint attention, understanding of desires and true beliefs) are evident earlier in development (Charman, Baron-Cohen, Swettenham, Baird, Cox, & Drew, 2000; Wellman & Liu, 2004; Wellman & Woolley, 1990) and that ToM continues to improve throughout childhood and beyond (Happé, 1994; Kaland, Møller-Nielsen, Smith, Lykke Mortensen, Callesen, & Gottlieb, 2005; Liddle & Nettle, 2006).

A large body of research has found that children and adolescents with autism spectrum conditions (ASC) experience difficulties understanding mental states or show delayed ToM development (Baron-Cohen, 2000) and its emergence is often linked to increased verbal skills (Lockett, Powell, Messer, Thornton, & Schultz, 2002). One of the core diagnostic criterion linked to ASC is a difficulty in reciprocal social interaction and communication (American Psychiatric Association, 2000). In typical development, better ToM skills have been linked to more effective and extensive social relationships in both children (Liddle & Nettle, 2006; McElwain & Volling, 2004) and adults (Stillier & Dunbar, 2007). The relationship between ToM and social behavior has also been supported in some studies with children and adolescents with ASC (Frith, Happé, & Siddons, 1994); although further research has not shown this link (Plumet & Tardiff, 2005). Researchers increasingly recognize that ToM deficits are likely to be only one of several elements that can explain the profile of social and communication difficulties in autism (Tager-Flusberg, 2007). Given the positive associations between ToM and social behavior in typically developing children and, in

some individuals with ASC, theoretical and empirical work has continued to investigate ToM as one factor that potentially underpins social behavior in autism and several studies have developed methods to teach ToM to this population.

The majority of these studies have aimed to establish whether it is possible to show improved understanding in ToM through intensive teaching. Teaching effectiveness has been assessed by exploring its impact on structurally and conceptually similar and dissimilar tasks. Researchers have suggested that a demonstration of teaching effects to non-taught tasks is critical to argue that any conceptual change has taken place (Iao, Leekam, Perner, & McConachie, 2011; Knoll & Charman, 2000). Further research has assessed the broader impact of ToM teaching on related social and communication skills. In addition, given the recognized difficulties in generalizing ToM to social behavior (Frith & Happé, 1994), a small number of studies have included social skills training in conjunction with ToM teaching in order to help children and adolescents with ASC to understand the relevance of mental states in day-to-day situations.

In this review, we highlight the diverse and innovative methods that have been developed over the last 15 years to foster the development of mental state understanding in children and adolescents with ASC. We present studies that have focused on teaching false belief understanding, as well as those that have adopted a developmental approach to teach related constructs that, in typical development, have been found to emerge before or after false belief understanding. The overall aim of the chapter is to assess the effectiveness of teaching ToM to children with ASC and to consider possible future directions for the development of this research. We specifically focus on studies that have aimed to teach children about beliefs as representations of the world and how they link to behavior and emotion. (See the Appendix for a summary of teaching studies.) In addition, because the typical approach to teaching ToM requires children having basic language skills, we also explore a growing body of research that has aimed to teach non-verbal constructs (e.g. joint attention) related to the development of ToM.

Teaching individuals with autism spectrum conditions to understand false belief

In a typical ToM scenario, an object is transferred from one location to another without the protagonist's knowledge, so that this character then has a false belief about its location (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1993). Children are asked to predict where a protagonist will go to retrieve the object or where he or she thinks the object is (reviews by Doherty, 2009; Wellman et al., 2001). Young children and children with ASC often make behavioral and emotional predictions based on the objects real location, and not on the protagonist's belief about where or what the object is (Baron-Cohen et al., 1985). Variations of this traditional transfer task have been developed to assess conceptually equivalent constructs: children's understanding that people can hold false beliefs about the contents of a container (i.e. the deceptive appearance task) or the difference between what an object really is vs. what it appears to be (the appearance-reality distinction; Flavell, Flavell, & Green, 1983). Further studies have measured children's understanding of how a protagonist feels when they discover that their belief is false (Hadwin & Perner, 1991; Harris, Johnson, Hutton, Andrews, & Cooke, 1989). Others have looked at the behavioral or emotional consequences around embedded beliefs (i.e. beliefs about beliefs or second-order ToM; Perner & Wimmer, 1985). Moving beyond the false belief literature Happé (1994) also developed a set of strange stories to explore children's understanding of ToM-related social constructs, such as sarcasm, irony, and humor.

Teaching ToM to children and adolescents with ASC has to some extent captured the diversity of this basic literature. Wellman and colleagues, for example, taught children with ASC to

understand false beliefs via transfer task using thought bubbles (Wellman, Baron-Cohen, Caswell, Carlos Gomez, Swettenham, Toye, et al., 2002). Children were taught in five stages, from introducing thought bubbles (stage 1) to working through increasingly complex tasks to highlight a protagonist's thought about an object's location, to understanding false belief (stage 5). After teaching, children were asked to think about a character's thoughts without the use of a thought bubble. Across two studies, the results consistently showed that the majority of children learnt to pass false belief tasks and they were able to transfer that knowledge to similar tasks. When tested on novel paradigms that were conceptually easier or equivalent to taught tasks the results were less clear: only a quarter of the children showed some ability to pass these tasks across the two studies. However, because most children were able to progress through the learning stages, the authors argued that the use of thought bubbles represents an effective and simple strategy to teach children with ASC to understand mental states.

Other researchers have used photographs to teach mental state understanding to children with ASC. Swettenham and colleagues used photographs to teach ToM based on previous work, which found that children with ASC understand that photographs represent events in the world (Swettenham, Baron-Cohen, Gomez, & Walsh, 1996). The study used a manikin's head with a slot in the top where pictures could be placed to convey the idea of mental states as representations of the world. Children with ASC were taught to think about false beliefs for 1 hour a day over 5 days in four steps, which focused initially on the picture-in-the-head analogy (step 1), and progressed to consider links between this analogy and mental states or behaviors (steps 3 and 4). The study provided corrective feedback and explanations for incorrect answers. It showed that most children were able to understand that what people see can be represented as a photo in the head and that this representation was linked to subsequent behavior. In addition, it showed that links between photos and behavior generalized more broadly to show some understanding of knowledge acquisition (seeing leads to knowing) and false belief, as measured by the deceptive appearance task. The authors argued that successful demonstration of passing novel tasks in some children supported the use of photographs to teach ToM to children with ASC.

McGregor, Whiten, & Blackburn (1998) adopted this approach to teach false belief understanding to children and adults with ASC and typically developing children compared with a non-teaching control group. The study used errorless learning techniques across three teaching schedules. The authors argued that a minimum amount of correction is most likely to promote conceptual learning. Initially, the authors minimized verbal input and increased the use of visual cues by emphasizing a doll protagonist's intentions in a false belief transfer task. In the second schedule, the authors used the picture-in-the-head technique and the third combined both of these teaching methods. The results showed that performance in passing false beliefs tasks in both typically developing children and children with ASC was significantly better in the experimental compared with the control group; indicating that all children and adults were able to learn to pass ToM tasks. Further analysis also revealed that learning was most effective in both groups using the picture-in-the-head technique. In addition, both teaching groups were able to use the instruction to make correct novel predictions about their own false beliefs, and typically developing children also showed some generalization from teaching to tasks that involved false beliefs judgments based on real life actors. McGregor et al. suggested that, because participants were able to use teaching to pass non-taught tasks, some conceptual change must have taken place. They proposed that future research should use the picture-in-the-head approach combined with errorless learning techniques to further assess the impact of change in individuals with ASC to real life settings.

The photograph method has also been used in a study that taught children with ASC to understand mental states, as well as to develop their executive function (EF) skills (Fisher & Happé, 2005).

EF (e.g. planning, inhibition, flexible thinking) and mental state understanding are associated in typical development (Pellicano, 2007). Specific aspects of EF (e.g. monitoring) have been found to be impaired in children with ASC (Happé, Booth, Charlton, & Hughes, 2006). In this teaching study, children were randomly allocated to groups that focused on either EF or ToM. Teaching in both domains was delivered in five stages and to criterion, so that the overall amount of teaching ranged between 4 and 10 days, with 25 minutes per day. EF teaching focused on the Wisconsin card-sorting task reflecting mental flexibility. This task requires children to initially sort or match cards according to one dimension (e.g. colour). Each time a child sorts a card they are told if they are right or wrong. During the task the sorting rule changes (e.g. from colour to number) and the time children take to change the rule as indicated in the number of errors they make is used as a marker of performance. Teaching involved the use of “brain tools” (laminated pieces of card used to demonstrate different sorting dimensions) to complete tasks that require flexibility of thought (stage 1) to independent flexibility at stage 5. ToM teaching also involved five stages based on the picture-in-the-head method that used dolls with slots in their heads so photographs could be inserted. It ranged from teaching children that thoughts are like pictures in the head (stage 1) to understanding thoughts without pictures (stage 5).

The results showed that children were able to move through the ToM stages with teaching, with the majority of children showing improvements after teaching and at follow up. Teaching also led to improved performance on a seeing leads to knowing task. Interestingly, the study showed that improvement in mental state understanding was positively associated with language, indicating that children with better language skills might benefit most from ToM teaching. Children also improved on EF tasks through teaching, although there was no evidence of generalization to non-taught EF tasks. Teaching EF also led to improvement in ToM, where this effect was most significant at follow-up. The authors suggested that improvements in ToM via EF teaching might reflect the similarity between ToM and EF tasks (both require flexibility of thought and shifting between rules or different states of the world). Importantly, ToM allows individuals to reflect on their own plans and intentions or what they know and do not know. While children in both teaching groups showed improvement in EF and ToM, this change was not reflected in day-to-day behavior as reported by teachers.

The use of structured progressive techniques and repetition to teach ToM understanding is reflected in early teaching studies. Swettenham (1996) used computers to teach false belief. He suggested that children with ASC would benefit from a computer medium because they could control their speed of learning in a predictable environment. The study compared the effectiveness of teaching in children with ASC, typically developing and children with Down syndrome. Teaching consisted of two short sessions over 4 days and included a 3-month follow-up assessment. The results showed that over eight teaching sessions, all groups showed learning. The typically developing children and children with ASC did not differ in their learning rate and both groups were significantly better learners compared with the group of children with Down syndrome. All children were able to pass non-taught tasks at the follow-up session. In addition, the majority of typically developing children and those with Down syndrome were also able to pass novel ToM tasks immediately after teaching and this learning was still evident in the typical group at follow-up.

LeBlanc and colleagues used technology via video modeling and additional cues to teach perspective taking to three children with ASC (LeBlanc, Coates, & Daneshvar, 2003). Children were assessed on a standard false belief task, a hide and seek task (a transfer task that used footprints as behavior cues to locate hidden objects) and a deceptive appearance task. Videos were developed to show the correct answers to the hide and seek and deceptive appearance tasks and rewards (e.g. food, stickers) were given for correct answers. The results indicated that all three children who took

part were able to pass similar ToM tasks as a result of teaching. In addition, two of the three children also passed a novel false belief task after teaching. The authors argued that a video medium is useful for ensuring that children attend to relevant cues to understand the perspective of another. In addition, they suggested that this approach to teaching serves to enhance motivation and attention in this group of children.

Moving on from videos, Bowler & Strom (1998) used actors to teach children with ASC, typically developing children, and children with learning difficulties false belief using four different versions of a transfer task. Each version provided increasingly salient cues to the actors's belief, where these included no cue (standard task); the person going to the original location where the object was hidden (behavioral cue); the person looking surprised when the object was not there (emotional cue); and finally exposing participants to their own belief violations (own false belief cue). The study also included a control group of children who experienced the standard false belief transfer task (three times), as well as their own false belief cue. Like some previous research (McGregor et al., 1998), this study relied on the salience of the cues and no corrective feedback was given for incorrect responses. The study showed that children with ASC and typically developing older children were able to use the behavioral and emotional cues to pass a greater number of false belief questions compared with the control group. Younger typically developing children and those with learning difficulties did not benefit from these additional cues. The authors suggested that there is an age below which typically developing children are unable to use cues to pass false belief tasks. Consistent with Fischer & Happé (2005), they also suggested that false belief understanding depends on verbal ability; children with learning difficulties who had low verbal ability relative to the other groups made less progress.

While several studies have taught first order ToM using a false belief transfer task, one study used the understand the appearance–reality (A–R) distinction as the basis for teaching (Starr & Baine, 1996). A–R tasks involve assessing children's understanding that making outward and superficial changes to an object (e.g. making a white pencil look red by looking at it through a red filter) does not influence or alter its identity (see Doherty, 2009). Starr and Baine taught children with ASC to understand A–R size and colour distinctions through Direct Instruction (i.e. introduce the task and model the correct answer). Teaching consisted of working through colour and size tasks in two daily sessions for 5 days or until children were able to provide correct answers. The results showed that while some children were able to provide correct answers to the tasks during teaching and for non-taught similar A–R tasks after teaching, this learning was not maintained over time (see Swettenham, 1996 for similar results). The authors argued that their result suggests that active teaching underpinned children's ability to pass tasks.

A developmental approach to teaching theory of mind

Several have taught children with ASC to understand mental states by following the typical developmental stages leading up to false belief understanding (e.g. Begeer, Gevers, Clifford, Verhoeve, Kat, Hoddenbach, et al., 2011; Hadwin, Baron-Cohen, Howlin, & Hill, 1996; Howlin, Baron-Cohen, & Hadwin, 1999; Ozonoff & Miller, 1995). Hadwin and colleagues taught constructs linked to the development of ToM including pretend play and emotional understanding, as well as more basic perspective taking. Children with ASC received eight 30-minute sessions in one of these three teaching domains. Teaching in each domain was split into five levels; sessions started with developmentally more simple tasks and progressed onto more advanced tasks. For example, perspective taking started with very basic ideas (that we can see the same object in different ways) and moved through more complex tasks (seeing leads to knowing, true belief and

false belief.) Pretend play moved from sensorimotor play to imaginative play and emotion progressed from recognizing facial expression to belief-based emotion. In order to facilitate learning and generalization children were taught simple rules or principles that aimed to capture the conceptual level they were working at. The results showed that children were able to progress through the stages to show some improvement in their level of understanding. This change was most evident in the belief and emotion teaching domains where tasks were very structured. In addition, children showed some generalization to non-taught tasks and this learning was maintained at 2 months follow-up.

Part of this research also assessed the impact of teaching emotions, belief and pretend play on children's ability to expand conversation and on their use of mental state terms in speech (Hadwin, Baron-Cohen, Howlin, & Hill, 1997). Previous research had found that children with ASC show specific deficits in the pragmatic or social aspects of language. Tager-Flusberg (2000), for example, highlighted that conversational ability, including expanding on current conversational topics or adding new topics into a conversation, was poor in individuals with ASC and these pragmatic difficulties were attributed to deficits in ToM (review by Nilsen & Fecica, 2011). Hadwin and colleagues explored whether teaching ToM skills would enhance conversational skills in ASC and lead to an increased use of mental state terms in speech. They found that children's ability to expand and introduce new topics into a discussion was positively associated with children's expressive and receptive language abilities, but these abilities did not change as a result of ToM teaching. Further research has shown that teaching conversational skills to children with ASC does not improve ToM skills; Chin & Bernard-Opitz, 2000.

Ghim, Lee & Park (2001) similarly used developmental levels to teach perspective taking, knowledge acquisition and belief understanding to children with ASC and typically developing children. They showed improvement with teaching to pass non-taught and novel tasks, where this effect was maintained 2 weeks later. Children with learning difficulties did not show improvement equivalent to the other two groups. Similarly, McGregor, Whiten & Blackburn (1998) worked with adults and children with ASC and utilized developmental levels in teaching ToM. Teaching consisted of two or three 1-hour sessions taught at three levels of understanding. The first level focused on seeing-leads-to-knowing, the second on a story-based false belief and the third on video presented false belief scenarios. The use of mixed media was employed to facilitate transfer of knowledge to real life. This study used "errorless learning" techniques, such that children were able to see the belief of the story-based protagonist via pictures in their head (dolls with slots in their heads). The results showed that six of the ten participants were able to pass non-taught false beliefs tasks after teaching. Like other researchers (e.g. Swettenham et al., 1996), the authors suggested that the picture-in-the-head technique for teaching ToM was a useful tool to help individuals with ASC to understand that people can hold representations about the world.

While the majority of intervention studies have taught children individually, two recent reports utilized a social cognitive intervention developed by Steerneman and colleagues to teach ToM to small groups of children and adolescents with ASC (Begers, Gevers, Clifford, Verhoeve, Kat, Hoddenbach, & Boer, 2011; Gevers, Clifford, Mager & Boer, 2006; see Steerneman, Jackson, Pelzer & Muris, 1996). In both studies groups of 5–6 children with ASC were taught concepts related to ToM every week for around 3–4 months. Teaching focused on basic ToM skills, such as recognizing emotions, pretence and imitation, beliefs and false beliefs, and also included more advanced constructs reflecting second order tasks ToM, as well as irony and humor. In addition, parents were introduced to ToM and over five monthly sessions were taught methods to encourage their children to use ToM skills in everyday life. Both studies assessed ToM before and after teaching using a battery of ToM tasks to measure simple and more advanced ToM skills (see Muris, Steernman,

Meesters, Merckelbach, Horselenberg, van den Hogen, et al., 1999). In addition, parent report adaptive behavior in both studies was measured before and after teaching.

Gevers, Clifford, Mager, & Boer (2006) found that children with ASC showed improvements in imitation, as well as pretend play, humor and belief understanding. In addition, parents reported improved scores on some aspects of social adaptation related to interpersonal relationships, social skills and play. No detail on individual differences in learning was provided. Begeer et al. (2011) used the same intervention program in a study for a larger group of high functioning children with ASC who were randomly allocated to a teaching ToM or waiting list control group. As well as running sessions for parents, this study also involved parents in the last 15 minutes of every child intervention session. In addition, it measured the impact of ToM teaching on self-report emotional awareness and empathic skills. Overall, the intervention groups showed improvement on the overall ToM skills relative to the control group. The benefits of teaching in the ToM group, however, showed a mixed profile of learning. Children showed some improvement post-teaching in assessments related to belief understanding, but there was no overall improvement in the more basic or advanced ToM skills. In addition, the ToM group showed some improvement on tasks tapping an understanding of emotions; but there was no group difference for empathic or parent report social skills. The lack any improvement in ToM tasks via teaching to the use of these skills in real life settings is consistent with previous research (e.g. Fischer & Happé, 2005).

Silver & Oakes (2001) used developmental levels to focus exclusively on teaching emotional understanding to children with ASC. Half of the children received lessons as usual and the other half completed a computer emotion trainer task. Similar to previous research (Hadwin et al., 1996), emotional concepts were taught from simple to complex using multiple examples at each level (recognizing facial expressions or basing emotions on others' desires, beliefs or likes and dislikes of characters). The results of this study showed that all children in the training group improved in their emotional understanding, compared with those in the control group. In addition, the study showed some generalization of teaching to non-taught tasks, as well as to novel tasks that assessed children's understanding of ToM related social constructs (Happé, 1994). Like Swettenham (1996), the authors argued that the use of a computer program was engaging and motivating for children. They proposed that some of the concepts taught at the more complex levels might have facilitated generalization to novel tasks.

While the focus of interest in the current review is on studies that aimed to link emotional understanding to underlying beliefs and desires, it is worth noting one further study that adopted a multimedia approach to teaching emotions to children with ASC. Golan and colleagues, taught emotions using the Mindreading DVD, a computer-based program designed for children from age 4 to adulthood (Golan & Baron-Cohen, 2006). The DVD contains 412 emotions portrayed by actors and actresses of different ages, male and female, and of different ethnicities, and contains both computer games and an emotions library. After 10 weeks of using this DVD for a minimum of 2 hours per week, this study found that individuals with Asperger syndrome improved significantly in emotion recognition. More recently, this research group has developed a second DVD by entitled *The Transporters* (2010). This DVD contains a children's animation program aimed at pre-school and primary school age children on the autistic spectrum. The characters in the film are all vehicles (trains, trams, tractors, cable cars), but with human faces that display appropriate (though exaggerated) emotional expressions for the short stories that form each episode. Research based on this work has found that after watching the DVD for just 15 minutes per day for a 1 month period, children with ASC improved significantly in their emotion recognition ability relative to a control group of children with the same diagnosis who did not watch the DVD, where this ability generalized to unseen material (Golan, Ashwin Granader, McClintock, Day, Leggett, et al., 2010).

Teaching theory of mind and social skills

One of the earliest ToM papers taught children with ASC to understand both first order and more advanced ToM tasks, including constructs linked to intention and deception (Ozonoff & Miller, 1995). Teaching in this study was extensive, compared with other studies, and consisted of 14 sessions of 90 minutes that spanned over 4 months. Half of the sessions focused on teaching children social skills (expressing interest, reading non-verbal signals and emotional expression) and the other half on teaching ToM. Social skills support also included social outings, as well as parties and children were encouraged to use their perspective taking skills during these activities. ToM sessions aimed to teach the underpinnings of perspective taking including how people acquire information or knowledge (e.g. seeing-leads-to knowing), as well as first and second-order ToM. The results indicated that four out of five children in the teaching group improved in their understanding of mental states compared with the four control group children. However, parent and teacher ratings of social skills did not differ between the teaching and control groups either before or after teaching. The authors concluded that while they were able to teach children with ASC to pass ToM tasks, children did not appear to access these skills in every-day social situations.

A similar study combined ToM teaching with social skills training using a single case study for a child with ASC who demonstrated difficulties in emotional regulation (Feng, Ya-yu, Lo, Tsai, & Cartledge, 2008). Training consisted of assessment and explanation of ToM tasks, followed by role play (or maintenance sessions) with a small group of peers who had good levels of social skills. Training took place on a one-to-one basis four times per week and continued for around 10 weeks. The ToM assessment and teaching followed a developmental pattern, considering basic and more advanced understanding of desire and belief based emotion, as well as first and second order false beliefs, where these were embedded in situations the child would typically experience. The social skills training addressed difficulties with emotional regulation and appropriate expression of needs and these were assessed through observation. In contrast to previous research (Ozonoff & Miller, 1995; Begeers et al., 2011), the results showed improvement in ToM skills, as well as in social interaction. In addition, the authors highlighted that these skills generalized to novel settings. The authors attributed this positive result to the combined effect of ToM and social skills training, the use of maintenance sessions and the varied teaching approach adopted during training (i.e. the use of multiple and dynamic exemplars, role play and the inclusion of maintenance sessions). Though the results of this study are promising, this basic approach needs to be extended to confirm the finding in larger groups of children and with those with varying abilities.

Joint attention and theory of mind

Linked to a developmental approach to understanding and teaching ToM in children and adolescents with ASC, one further construct that has been highlighted as an important precursor to its development is joint attention (Howlin, 2008). Joint attention (JA) refers to the ability to coordinate attention with social partners to objects and events in the environment, using gaze, gesture or language (Scaife & Bruner, 1975). Typically, it emerges between the ages of 8 and 15 months (Bakeman & Adamson, 1984; Jones, Carr, & Feeley, 2006) and researchers have argued that its development is critical for early social communication (Bakeman & Adamson, 1984). JA skills have been separated into two categories—responding to JA (reflecting an ability to follow the line of regard and points of others); and initiating joint attention (which includes making eye contact, pointing and showing items of interest in order to share an item or event with another person; see reviews by Meindl & Cannella-Maone, 2011; Mundy & Jarrold, 2010).

Disruptions in JA processes (e.g. a lack of interest in parents' gaze, reduced eye contact, failure to point for interest) are often noted in young children with ASC (Charman, Baird, Simonoff, Loucas, Chandler, Meldrum, et al., 2007; 2003; Mundy, Sigman, & Kasari, 1990) and this absence is considered to be one of the first indices of social and communication difficulties in this population (Murray, Craghead, Manning-Courtney, Shear, Bean, & Prendeville, 2008; Whalen, Schriebman, & Ingersoll, 2006). Early home recordings of children who later went on to receive a diagnosis of ASC have highlighted a lack of JA skills. For example, two studies examined first year birthday party videotapes to look for the presence or absence of early JA (Osterling & Dawson, 1994; Werner, Dawson, Osterling, & Dinno, 2000). At 12 months of age children who went on to receive a diagnosis of ASC (compared with typically developing peers) were less likely to look at others, to show or point at objects, or respond to their name being called. In a further study, failure to initiate JA (pointing for interest) or to engage in pretend play at 18 months of age was linked to a later diagnosis of ASC in 80% of cases (Baron-Cohen, Cox, Baird, et al., 1996). Difficulties responding to JA at 14 months was also shown to be predictive of an ASC diagnosis by two years of age (Sullivan, Finelli, Marvin, Garrett-Mayer, Bauman, & Landa, 2007). In a related line of research, younger siblings of children with ASC were monitored over time to see whether any early behavior would predict a later diagnosis of ASC. The results highlighted that all of the children who went on to receive a diagnosis showed early social communication impairments, characterized as reduced interest or pleasure in others, and fewer initiations of social interaction (limited eye contact and social smiling, and a lack of pointing to or sharing items of interest with others; Bryson, Zwaigenbaum, Brian, Roberts, Szatmari, Rombough, et al., 2007).

While some theories suggest that JA and ToM reflect a common pathway of atypical development in children with ASC, research has generally lagged behind in linking up these two constructs. One study measured the extent to which infants at 20 months of age spontaneously switched their attention or looked between a moving toy and another person. They found that the presence of these JA skills at 20 months was linked to ToM ability when infants were 44 months of age; even when initial IQ and language skills were taken into account. The study supports the proposition that JA is a developmental precursor to ToM (Charman et al., 2000). Similarly, Charman (2003) found that better JA skills at 20 months were associated with improved language outcomes and less symptom severity at 42 months in a sample of children with ASC.

ToM intervention studies that have adopted a developmental approach have tended to focus on teaching tasks related to the origins of knowledge and beliefs (e.g. Begeer et al., 2011). The last 15 years has also seen a proliferation of research studies focusing on teaching JA skills to children with autism. While this research largely sits beyond the scope of the current review, there are some key findings and similarities between these two literatures that are important to note. Some intervention studies have aimed to teach or focus on aspects of JA attention that emerge earlier in development or that are evident in children with ASC. Some evidence suggests that the different components of JA dissociate in development. Responding to JA, for example, is not problematic for children with ASC who have a mental age over 30–36 months, while initiating JA is typically absent in children with ASC, irrespective of age (Mundy, Sigman, & Kasari, 1994; review by Mundy & Jarrold, 2010). Other researchers have argued that responding to JA is a developmental precursor to initiating JA and that these two skills should be addressed sequentially and separately in intervention programs (Kasari, Gulsrud, Wong, Kwon, & Locke, 2010; Murray et al., 2008). However, similar to ToM studies, some difficulties with generalization have been found; while several studies have been able to show an improvement in responding to JA in young children with ASC the emergence of this skill has typically not led to any change in initiating JA (Schertz & Odom, 2007; Taylor & Hoch, 2008).

Similar to ToM interventions, several different methods have been used to teach children with ASC to initiate or respond to JA (or both) including those that adopt highly structured applied behavior analysis techniques (Taylor & Hoch, 2008) or others that have used more naturalistic play routines (Kasari, Freeman, & Paparella, 2006). In addition, they have used different agents to facilitate change (researchers, parents, peers). Kasari et al., (2010), for example, used caregivers of children with ASC to encourage generalization of skills outside the JA teaching sessions. They randomized 38 caregiver and toddler pairs (mean age = 30.82 months) into intervention and waiting list control groups, and taught responding JA followed by instruction in initiating JA. Play routines were used whereby the parent was taught to follow a child's interest, and expand their play and joint attention repertoire. Relative to the waiting list control group, they found that children in the intervention group showed more joint engagement and specifically increased responding to JA. In contrast, however, initiating JA was not found to improve, even though it was directly targeted by parents.

One further issue in this literature relates to the consequences for children who successfully engage in JA during teaching or training. Within a teaching paradigm, consequences for successful joint engagement can be non-social (e.g. a tangible item) or social (e.g. social attention). The function of JA in typical development is social communication and attention. Some researchers have suggested that intervention programs should aim to provide naturalistic social attention as a consequence to a child engaging in joint attention (review by Meindl & Cannella-Malone, 2011). Two recent studies have successfully taught children with ASC to engage in JA using social attention as a consequence. Naoi, Tsuchiya, Yamamoto, & Nakamura (2008), for example, trained three children between the ages of 5 and 8 years to initiate JA using preferred items as the attending stimuli and adult attention as the social consequence and found that initiating JA increased for all three participants. Taylor & Hoch (2008) also reported positive effects of training both responding and initiating JA in three children aged 3–8 years using social interaction and physical contact as consequences for engaging in joint attention; where this strategy was successful for two of the three children. Similar to Feng et al. (2008), the results serve to highlight that embedding teaching in a social context has beneficial effects for learning and generalization.

Summary

Most studies have found that children and adolescents with ASC are able to learn to pass ToM tasks through teaching and to transfer their knowledge to conceptually similar tasks that had not been directly taught (e.g. Beger et al., 2011; Bowler & Strom, 1998; Wellman et al., 2001). Further studies have shown that some children were able to generalize learning to novel tasks (Fischer & Happé, 2005; LeBlanc Coates, Daneshvar, Charlop-Christy, & Morris, 2003; Silver & Oakes, 2001) and researchers have argued that this type of generalization does reflect conceptual change. In general, studies have used a structured approach to teaching ToM that allows children to practice tasks through the use of multiple examples and some researchers have argued this approach to teaching is the most beneficial for learning (Hadwin et al., 1996; Silver & Oakes, 2001; Swettenham, 1996; see also Iao, Leekam, Perner, & McConachie, 2011). Other researchers have suggested that a developmental approach can facilitate the development of ToM, where teaching can mirror learning that occurs in typical development. In general, this approach has found that children can move from basic to more complex constructs (Hadwin et al., 1996, 1997; Silver & Oakes, 2001). Consistent with the emergence of ToM skills in children and adolescents with ASC (Luckett, 2002), the results of some studies indicate that the benefits of ToM teaching are moderated by language ability, with those children who have better language skills showing most change (Bowler & Stromm, 1998;

Fischer & Happé, 2005). Further research should explore more clearly why some children and adolescents with ASC are able to learn ToM skills through teaching, while others show little benefit.

The benefits of teaching ToM to broader social and communication skills is mixed: some studies have been able to demonstrate a positive impact of ToM teaching to improved social behavior more generally (Gevers et al., 2006; Feng et al., 2008). Considering results more broadly, however, this result is uncommon (Beeger et al., 2011; Fischer & Happé, 2005; Hadwin et al., 1997; Ozonoff & Miller, 1995). These findings reflect the mixed picture in basic research in children with ASC that has considered associations between ToM understanding and its use in social interaction and communication (review by Nilsen & Fecica, 2011). This lack of generalization has raised questions about what children learn during ToM teaching; do they learn about mental states or have they acquired strategies for passing tasks that they experience repeatedly and that have a reliable and predictive structure and format (Swettenham, 2000; Howlin, 2008). The lack of generalization of ToM to day-to-day behavior has led some researchers to argue that longer term ToM teaching built into an educational curriculum, or a combination of ToM teaching alongside social skills workshops, would be most beneficial in terms of developing and utilizing ToM skills in different contexts (Feng et al., 2008; Ozonoff & Miller, 1995). In order to more clearly understand different formats in teaching ToM, future research needs to adopt a more rigorous randomized control design approach in the development of intervention studies to enable a better understanding of the benefits of ToM (Beeger et al., 2011). In addition, studies that are able to consider the impact of change at a neural level (e.g., to demonstrate increased activity in brain regions known to be activated during theory of mind tasks) would allow researchers to more clearly demonstrate the impact of teaching on understanding of key constructs (review by Carrington & Bailey, 2009).

The majority of studies start teaching first-order ToM; around what a typical 3- or 4-year-old child would understand. Some researchers have proposed that interventions can be targeted at recognized precursors of ToM, such as pretend play or joint attention (Howlin, 2008). Where this approach to intervention has emphasized social reward and social application children have been found to benefit more from teaching (e.g. Taylor & Hoch, 2008). However, similar to the ToM intervention research, this literature is also limited by the number of studies that have explored the extent to which JA skills are generalized outside of the training setting, and whether the social function of JA is effectively established and maintained over time. Interventions targeted toward increasing JA skills may serve to scaffold or facilitate the development of social interaction and language ability, which may, in turn, have a positive impact on the development of ToM. One benefit of teaching JA is that its impact on children with less language ability can be explored and further research should assess the relationship between JA teaching with language development and the emergence of basic ToM constructs (see Charman et al., 2000).

Despite some of the limitations in this literature, many authors have acknowledged that teaching ToM can be useful to individuals with ASC in order to help them to think about mental states; something that they would not do naturally in the course of development. The main aim of any intervention is to give children and adults the basic tools to understand mental states in order to help them negotiate their social world.

Acknowledgements

The writing of this chapter was supported by an Action Medical Research grant awarded to the first author (SP4598): the leading UK-wide medical research charity dedicated to helping babies and children.

References

- American Psychiatric Association (2000). *Diagnostic and Statistical Manual for Mental Disorders*, 4th edn. Washing DC: American Psychiatric Association.
- Bakeman, R. & Adamson, L. B. (1984). Coordinating attention to people and objects in mother–infant and peer–infant interaction. *Child Development* 55: 1278–89.
- Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. In: S. Baron-Cohen, H. Tager-Flusberg, D. J. Cohen (Eds), *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*, 2nd edn (pp. 3–20). New York: Oxford University Press.
- Baron-Cohen, S., Baldwin, D., & Crowson, M. (1997). Do children with autism use the Speaker's Direction of Gaze (SDG) strategy to crack the code of language? *Child Development* 68: 48–57.
- Baron-Cohen, S., Cox, A., Baird, G., Swettenham, J., & Nighingale, N. (1996). Psychological markers in the detection of autism in infancy in a large population. *British Journal of Psychiatry* 168(2): 158–63.
- Baron-Cohen, S., Leslie, A. M. & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition* 21: 37–46.
- Begeer, S., Gevers, C., Clifford, P., Verhoeve, M., Kat, K., Hoddenbach, E., & Boer, F. (2011). Theory of mind training in children with autism: A randomized controlled trial. *Journal of Autism and Developmental Disorders* 4: 997–1006.
- Bowler, D. M., & Strom, E. (1998). Elicitation of first-order theory of mind in children with autism. *Autism* 2: 33–44.
- Bryson, S. E., Zwaigenbaum, L., Brian, J., Roberts, W., Szatmari, P., Rombough, V., & McDermott, C. (2007). A prospective case series of high-risk infants who developed autism. *Journal of Autism and Developmental Disorders* 37: 12–24.
- Carrington, S. J. & Bailey, A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping* 30: 2313–35.
- Charman, T. (2003). Screening and surveillance for autism spectrum disorder in research and practice. *Early Child Development and Care* 173(4): 363–74.
- Charman, T., Baird, G., Simonoff, E., Loucas, T., Chandler, S., Meldrum, D., & Pickles, A. (2007). Efficacy of three screening instruments in the identification of autistic-spectrum disorders. *British Journal of Psychiatry* 191: 554–9.
- Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Cox, A., & Drew, A. (2000). Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development* 15: 481–98.
- Chin, H. Y., & Berbard-Opitz, V. (2000). Teaching conversational skills to children with autism: Effect on the development of a theory of mind. *Journal of Autism and Developmental Disorders* 30: 569–83.
- Doherty, M. J. (2009). *Theory of Mind*. Hove: Psychology Press.
- Dunn, J. R., Brown, C., 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: 1352–66.
- Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Transitional period in the development of the appearance–reality distinction. *International Journal of Behavioural Development* 12: 509–26.
- Feng, H., Ya-yu, L., Tsai, S., & Cartledge, G. (2008). The effects of theory of mind and social skill training on the social competence of a sixth grade student. *Journal of Positive Behaviour Intervention* 10: 228–42.
- Fisher, N., & Happé, F. (2005). A training study of theory of mind and executive function in children with autistic spectrum disorders. *Journal of Autism and Developmental Disorders* 35: 757–71.
- Frith, U., Happé, F., & Siddons, F. (1994). Autism and theory of mind in everyday life. *Social Development* 3: 108–24.

- Gevers, C., Clifford, P., Mager, M., & Boer, F. (2006). Brief report: A theory of mind based social cognition training program for school-aged children with pervasive developmental disorders; an open study of its effectiveness. *Journal of Autism and Developmental Disorders* 36: 561–71.
- Ghim, H-R., Lee, H., & Park, S. (2001). Autistic children's understanding of false belief: studies based on computerized animation task. Proceedings of the First International Workshop on Epigenetic Robotics, Sweden.
- Golan, O., Ashwin, E., Granader, Y., McClintock, S., Day, K., Leggett, V., & Baron-Cohen, S. (2010). Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotion faces. *Journal of Autism and developmental Disorders* 40: 269–79.
- Golan, O., & Baron-Cohen, S. (2006). Systemizing empathy: teaching adults with Asperger syndrome of high-functioning autism to recognize complex emotions using interactive media. *Development and Psychopathology* 18: 591–617.
- Hadwin, J. A., Baron-Cohen, S., Howlin, P., & Hill, K. (1996). Can we teach children with autism to understand emotions, belief or pretence? *Development and Psychopathology* 8: 345–65.
- Hadwin, J. A., Baron-Cohen, S., Howlin, P., & Hill, K. (1997). Does teaching a theory of mind have an effect on social communication in children with autism? *Journal of Autism and Developmental Disorders* 27: 519–37.
- Hadwin, J. A., & Perner, J. (1991). Pleased and surprised: children's cognitive theory of emotion. *British Journal of Developmental Psychology* 9: 215–34.
- Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders* 24: 129–54.
- Happé, F., Booth, R., Charlton, R., & Hughes, C. (2006). Executive function deficits in autism spectrum disorders and attention-deficit/hyperactivity disorder: Examining profiles across domains and ages. *Brain and Cognition* 61: 25–39.
- Harris, P. L., Johnson, C. N., Hutton, D., Andrews, G. M., & Cooke, T. (1989). Young children's theory of mind and emotion. *Cognition and Emotion* 3: 379–400.
- Howlin, P. (2008). Can children with autism spectrum disorders be helped to acquire a theory of mind? *Revista de Logopedia, Foniatria y Audiologia* 28: 74–89.
- Howlin, P., Baron-Cohen, S., & Hadwin, J. A. (1999). *Teaching Children with Autism to Mind-read*. Chichester: John Wiley and Sons.
- Iao, L., Leekam, S., Perner, S., & McConachie, H. (2011). Further evidence for non-specificity of theory of mind in preschoolers: Training and transferability in the understanding of false beliefs and false-signs. *Journal of Cognition and Development* 12: 56–79.
- Jones, E. A., Carr, E. G., & Feeley, K. M. (2006). Multiple effects of joint attention intervention for children with autism. *Behavior Modification* 30: 782–834.
- Kaland, N., Møller-Nielsen, A., Smith, L., Lykke Mortensen, E., Callesen, K., & Gottlieb, D. (2005). The Strange Stories test: A replication study of children and adolescents with Asperger syndrome. *European Child and Adolescent Psychiatry* 14: 73–82.
- Kasari, C., Freeman, S., & Paparella, T. (2006). Joint attention and symbolic play in young children with autism: a randomized controlled intervention study. *Journal of Child Psychology and Psychiatry* 47(6): 611–20.
- Kasari, C., Gulsrud, A. C., Wong, C., Kwon, S., & Locke, J. (2010). Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism. *Journal of Autism and Developmental Disorders* 40: 1045–56.
- Knoll, M., & Charman, T. (2000). Teaching false belief and visual perspective taking skills in your children: can a theory of mind be trained? *Child Study Journal* 30: 273–304.
- LeBlanc, L. A., Coates, A. M., Daneshvar, S., Charlop-Christy, M. H., & Morris, C. (2003). Using video modeling and reinforcement to teach perspective-taking skills to children with autism. *Journal of Applied Behaviour Analysis* 36: 253–7.

- Liddle, B., & Nettle, D. (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology* 4: 231–46.
- Luckett, T., Powell, S. D., Messer, D. J., Thornton, M. E., & Schultz, J. (2002). Do children with autism who pass false belief tasks understand the mind as active interpreter? *Journal of Autism and Developmental Disorders* 32: 127–40.
- McElwain, N. L., & Volling, B. L. (2004). Attachment security and parental sensitivity during infancy: Associations with friendship quality and false belief understanding at age four. *Journal of Social and Personal Relationships* 21: 639–67.
- McGregor, E., Whiten, A., & Blackburn, P. (1998). Teaching theory of mind by highlighting intention and illustrating thoughts: A comparison of their effectiveness with 3-year-olds and autistic individuals. *British Journal of Developmental Psychology* 16: 281–300.
- McGregor, E., Whiten, A., & Blackburn, P. (1998). Transfer of the picture-in-the-head analogy to natural context to aid false belief understanding in autism. *Autism* 2: 367–87.
- Meindl, J. N., & Cannella-Malone, H. I. (2011). Initiating and responding to joint attention bids in children with autism: A review of the literature. *Research in Developmental Disabilities* 32: 1441–54.
- Mundy, P., Sigman, M., & Kasari, C. (1994). Joint attention, developmental level, and symptom presentation in autism. *Development and Psychopathology* 6: 389–401.
- Mundy, P., Sigman, M., & Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and Developmental Disorders* 20: 115–28.
- Mundy, P., & Jarrold, W. (2010). Infant joint attention, neural networks and social cognition. *Neural Networks* 23: 985–97.
- Muris, P., Steernman, P., Meesters, C., Merckelbach, H., Horselenberg, R., van den Hogen, T., & van Dongen, L. (1999). The TOM test: A new instrument for assessing theory of mind in normal children and children with pervasive developmental disorders. *Journal of Autism and Developmental Disorders* 29: 67–80.
- Murray, D. S., Craghead, N. A., Manning-Courtney, P., Shear, P. K., Bean, J., & Prendeville, J-A. (2008). The relationship between joint attention and language in children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities* 23: 5–14.
- Naoi, N., Tsuchiya, R., Yamamoto, J-I., & Nakamura, K. (2008). Functional training for initiating joint attention in children with autism. *Research in Developmental Disabilities* 29: 595–609.
- Nilsen, E. S., & Fecica, A. M. (2011). A model of communicative perspective-taking for typical and atypical populations of children. *Developmental Review* 32: 55–78.
- Naoi, N., Tsuchiya, R., Yamamoto, J-I., & Nakamura, K. (2008). Functional training for initiating joint attention in children with autism. *Research in Developmental Disabilities* 29: 595–609.
- Osterling J., & Dawson, G. (1994). Early recognition of children with autism: A study of first birthday home videotapes. *Journal of Autism and Developmental Disorders* 17: 247–57.
- Ozonoff, S., & Miller, J. N. (1995). Teaching theory of mind: a new approach to social skills training for individuals with autism. *Journal of Autism and Developmental Disorders* 25: 415–33.
- Pellicano, E. (2007). Links theory of mind and executive function in young children with autism: clues to developmental primacy. *Developmental Psychology* 43: 974–90.
- Perner, J. (1993) *Understanding the Representational Mind*. London: MIT Press.
- Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: You catch it from your sibs. *Child Development* 4: 1228–38.
- Perner, J., & Wimmer, H. (1985). “John thinks that Mary thinks that ...” Attribution of second-order beliefs by 5 to 10 year-old children. *Journal of Experimental Child Psychology* 39: 437–71.
- Plumet, M-H., & Tardiff, C. (2005). Understanding the functioning of social interaction with autistic children. In: L. Anolli, S. Duncan Jr, M. S. Magnusson, & G. Riva (Eds), *The Hidden Structure of Interaction: From Neurons to Culture Patterns*. Amsterdam: IOS Press.

- Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature* 253: 265–6.
- Schertz, H. H., & Odom, S. L. (2007). Promoting joint attention in toddlers with autism: A parent-mediated developmental model. *Journal of Autism and Developmental Disorders* 37: 1562–75.
- Silver, M., & Oakes, P. (2001). Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others. *Autism* 5: 299–316.
- Starr, E. M., & Baine, D. A. (1996). Theory of mind and children with autism: A direct instruction approach to teaching the colour and size appearance reality distinction. *Exceptionality Education Canada* 6: 69–88.
- Steerneman, P., Jackson, S., Pelzer, H., & Muris, P. (1996). Children with social handicaps: An intervention program using a Theory of Mind approach. *Clinical Child Psychology and Psychiatry* 1: 251–63.
- Stiller, J., & Dunbar, R. (2007). Perspective-taking and social network size in humans. *Social Networks* 29: 93–104.
- Sullivan, M., Finelli, J., Marvin, A., Garrett-Mayer, E., Bauman, M., & Landa, R. (2007). Response to joint attention in toddlers at risk for autism spectrum disorder: A prospective study. *Journal of Autism and Developmental Disorders* 37: 37–48.
- Swettenham, J. (1996). Can children with autism be taught to understand false beliefs using computers? *Journal of Child Psychology and Psychiatry* 37: 157–65.
- Swettenham, J. (2000). Teaching theory of mind to individuals with autism. In: S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds), *Understanding Other Minds* (pp. 442–56). Oxford: Oxford University Press.
- Swettenham, J., Baron-Cohen, S., Gomez, J.-C., & Walsh, S. (1996). What's inside someone's head? Conceiving of the mind as a camera helps children with autism acquire an alternative to a theory of mind. *Cognitive Neuropsychiatry* 1: 73–88.
- Tager-Flusberg, H. (2007). Evaluating the theory-of-mind hypothesis of autism. *Current Directions in Psychological Science* 16: 311–15.
- Tager-Flusberg, H. (2000). What language reveals about the understanding of mind in children with autism. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds), *Understanding Other Minds* (pp. 124–9). Oxford: Oxford University Press.
- Taylor, B. A., & Hoch, H. (2008). Teaching children with autism to respond to and initiate bids for joint attention. *Journal of Applied Behavior Analysis* 41: 377–91.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development* 72: 655–84.
- Wellman, H. M., Baron-Cohen, S., Caswell, R., Carlos Gomez, J., Swettenham, J., Toye, E., & Lagattuta, K. (2002). Thought-bubbles help children with autism acquire an alternative to a theory of mind. *Autism* 6: 343–63.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72(3): 655–84.
- Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development* 75: 523–41.
- Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition* 35: 245–75.
- Werner, E., Dawson, G., Osterling, J., & Dinno, N. (2000). Brief report: Recognition of autism spectrum disorder before one year of age: A retrospective study based on home videotapes. *Journal of Autism and Developmental Disorders* 30: 157–62.
- Whalen, C., Schreibman, L., & Ingersoll, B. (2006). The collateral effects of joint attention training on social initiations, positive affect, imitation, and spontaneous speech for young children with autism. *Journal of Autism and Developmental Disorders* 36(5): 655–64.
- Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13: 41–68.

This page intentionally left blank

Section 4

Comparative and philosophical perspectives

This page intentionally left blank

Culture and the evolution of interconnected minds

Andrew Whiten

Introduction: The “deep social mind” of *Homo sapiens*

The now vast research literature on “understanding other minds” shows us that humans are inveterate mentalists, and studies of child development trace the construction of increasingly sophisticated cognitive penetration of the workings of the mind in self and others. As reviewed below, some non-human animals show some of the more elementary foundations of this capacity, but in its adult forms our baroque human mental interpenetration is unparalleled in its complexity and depth.

Why are we like this? Much of the analysis of our understanding of other minds is relatively mechanistic and concerned with how the system works and how it gets constructed during ontogenetic development. The present volume largely reflects these preoccupations. Evolutionary and comparative analyses of the kind discussed in this chapter do also address these questions, but are also very much concerned with the adaptive function of the capacities: what part they play in the lives of the species studied, and why they have evolved in the forms they have. What part does “understanding other minds” play in the larger picture of the particular forms of animal life that our species displays, and why did evolution come up with such an extraordinary phenomenon?

This chapter begins by reprising and elaborating a little on an answer to this question that sets human mindreading (aka theory of mind, as elsewhere in this volume) in a larger functional framework. In the next section I outline this framework in terms of four major human socio-cognitive characteristics that together constitute what I have described, from a comparative perspective, as a distinctively “deep social mind,” adapted to the ecological niche that characterized the most recent phase of our evolutionary history (Whiten, 1999, 2006).

The human socio-cognitive niche

What kinds of lives have our ancestors lived over the past million years or so that shaped the modern human mind? Much archaeological evidence converges to describe a form of subsistence characterized as hunting and gathering, that gave way to agriculture only in the last 10 000 years or so—a mere blink on the scale of evolutionary timespans (Whiten, 1999, 2006). An impressive example of this evidence is sophisticated wooden hunting spears that have been dated to as much as 400 000 years ago (Thieme, 1997). During the last century many contemporary small-scale societies also subsisted by hunting and gathering, and anthropologists established that across many varying tropical regions, this typically involves animals being hunted and plant foods gathered, with much of the bounty brought back to a home base to be shared in egalitarian fashion amongst the members of the band (Marlowe, 2005).

The archaeological evidence suggests that this niche, unprecedented in apes, has had long standing significance in shaping the human mind. It was an extraordinary outcome to an evolutionary story that began with our early ancestors of a few million years ago being faced with major loss of forest cover in Africa and becoming bipedal apes, venturing into new and more open savannah-woodland habitats. In doing so, they were exposed to an impressive array of formidable predators including several large canine and feline species. It is an extraordinary fact about our evolutionary history that our ancestors not only survived these predator risks, but evolved over time into big game hunters themselves, competing with these “professional predators” so well that hunting became a mainstay of their foraging niche, unlike their earlier ancestors and their ape relatives that remained in the forest.

How was this possible? A promising hypothesis offered by Tooby & deVore (1987) was that evolving humans developed what these authors characterized as a “cognitive niche,” in which such intelligent innovations as hunting with weapons and traps allowed attacks on prey that could be just as successful—often more so—than those mounted by existing mammalian predators using their specialized morphological “weapons” of teeth, strength, speed, and claws. Moreover, the cognitive niche could evolve faster than the latter through the exercise of intelligence.

This analysis is highly plausible as far as it goes, but I have suggested that it fails to capture several major elements of the psychology that underwrites the success of the human hunter-gatherer way of life. Together these elements describe not just a cognitive niche, but crucially a sociocognitive niche, that I have referred to by the expression “deep social mind” (Whiten, 1999, 2006). This encompasses four principal, inter-dependent cognitive and behavioural clusters. These include “understanding other minds,” but this is only one, integral part of a broader complex of adaptations. I shall first outline these and later focus on their evolutionary origins. They take unique forms in our own species, but precursors of all of them have been identified through careful comparative studies (Whiten and Erdal, 2012).

Egalitarianism and cooperation

Analyses by David Erdal and myself (Erdal & Whiten, 1994, 1996) systematically confirmed that across 24 hunter-gatherer ethnographies, egalitarianism is a hallmark of hunter-gatherers. It is manifested in several different, interconnected ways, most notably in food sharing, with foodstuffs brought back to the camp shared out according to need, rather than initial ownership or even kinship, and also in a lack of formal leadership. These characteristics are not seen in our closest ape relatives, but they appear to be universally associated with a hunter-gatherer way of life in small bands numbering only about 20–50 people (Marlowe, 2005). Such a small and interdependent social world probably characterized much of our most recent evolutionary history, spanning hundreds of millennia until the emergence of agriculture.

Cooperation within a hunter-gather band also takes forms unprecedented among non-human primates. These include not only cooperation within a whole suite of hunting, gathering and other subsistence activities like camp-making, but cooperation that spans and integrates these. An important instance of the latter is the typical division of labour in which men are largely responsible for hunting and women for gathering, with the resources they collect brought back to the central camp and shared out. Here, there is also information-sharing such that both hunters and gatherers’ foraging the next day may be guided by what other members of the band have witnessed the previous day—such as active burrows, or fruiting bushes, in particular parts of the home range. All this means that a human hunting-gathering band acts much like a formidable “group-level predator” that can compete very successfully with other species whose foraging niches significantly overlap, such as social carnivores like lions.

The next three characteristics to be outlined play crucial roles in facilitating this capacity of human hunting-gathering bands to operate much like a single, well-coordinated “superorganism.” The societies of the super-altruistic social insects, like bees and ants, have also deservedly been called superorganisms. What is distinctive about the human style of superorganism is the sophisticated forms of social cognition that underlie its nature and competitiveness. These reinforce each other in ways outlined further below.

Mindreading (MR—aka theory of mind)

There appear to be few formal studies of mindreading in hunter-gatherer societies, although Avis and Harris long ago (1991) demonstrated the emergence of false belief attribution in Baka pygmy children at a similar age to that common in industrialized societies. Cross-cultural differences in MR have been identified (Lillard, 1998), but the existence of a common MR core can reasonably be seen as a basic human capacity (Shahaiean, Peterson, Slaughter, & Wellman, 2011; Wellman, Fang, & Peterson, 2011).

Among the consequences of this at the band level are that the minds of its members interpenetrate each other to depths unknown in other primates, facilitating the operation of the band as the uniquely coordinated, superorganismic “group-level predator” described above.

Culture

The cultures of hunter-gatherers may appear “simple” in some ways compared with those of agricultural and industrial societies. The material cultures of the latter have become gigantic of course, whereas nomadic hunter-gatherers can often carry all their goods on their backs when they intermittently shift campsites. However, the cultures of hunter-gatherers vastly surpass anything known in other primates and extend to phenomena that are again crucial to operating as a highly successful group-level predator, including language, social norms and rules, and technologies for gathering, hunting and trapping prey (Hewlett, Fouts, Boyette, & Hewlett, 2011; Lee & deVore, 1968).

What is particularly distinctive in human culture is its cumulative character: the ability to continuously build on the culturally inherited achievements of past generations to construct artifacts and actions that would be beyond any single individual to invent in their lifetimes (Tomasello, 1999). The range of weaponry that allowed evolving humans to become big game hunters alongside the competing big cats and canines of Africa is one important example, that Tooby and deVore had in mind in writing of the “cognitive niche.” However, the larger phenomenon of culture (including cumulative culture) is plainly a socio-cognitive achievement.

Language

Our unique linguistic form of communication performs many functions in a hunter-gatherer band, but in the context of the present theme, it facilitates the coordination and information-sharing that allows a band to operate so well as a group-level predator. For example, it facilitates band-level negotiation and planning of the strategies to be adopted by different hunting and gathering sub-groupings, and their ultimate integration.

An adaptive complex: the distinctive hominid context for “understanding other minds”

Each of the four main socio-cognitive characteristics outlined above contribute to the concept of “deep social mind” in different ways, and these reinforce each other to constitute an adaptive

complex underwriting the hunting gathering niche that characterized our evolutionary history. Perhaps most obviously, MR makes human minds “deeply social” through the ways in which each mind deeply penetrates others in the social band. The mutual interpenetrations this encompasses mean that one can legitimately consider the band as sharing *a* mind—a concept expressed in everyday language when we say that having discussed some issue, we are “of one mind” on the matter.

In the case of culture, there is a different, although related sense of deep social mind at stake: our minds are deeply social in that they are largely populated by all we absorb from the accumulated culture we are born into. The contents of what is absorbed vary around the world of course, as well as over historical time, but in each case the mind is deeply socially shaped by its cultural acquisitions, from technologies to religions and other customs.

Language, in turn, makes our minds deeply social because, for example, people spend much time telling others about what is “on their mind” or asking what others have “in mind.” In that respect, language can be thought of as one, particularly powerful, tool to allow mindreading (and its converse when we broadcast the contents of our mind to others—“mindwriting?”) to operate.

Finally, coming full circle in the set of characteristics outlined above, egalitarianism is associated with a deeply social mind insofar as individual goals are subjugated to socially-leveiling ones, and sophisticated forms of hunter-gatherer cooperation mean that individuals’ actions are designed specifically to interdigitate with others’, to greater effect.

The central import of all I have said so far is that MR is embedded in a larger, multi-stranded phenomenon of deep social mind, the significance of which can be discerned in the behavioural and ecological niches that mark out the evolution of our species over past millennia. My emphasis has been on what is distinctive in all this, in comparison to our closest primate relatives. However, a wealth of research discoveries in the last decade or two in particular allows us to make substantial reconstructions of the earlier evolutionary foundations that shaped the architecture of these human specialties. In the context of the present volume, two of them seem to beg special attention—mindreading, and culture.

Mindreading in present-day and ancestral primates

This chapter spends less time on this topic than on the one relating to culture. What non-human primates (henceforth “primates”) take into account in relation to the states of minds of others has been subject to several excellent recent reviews that the reader can be referred to (Call & Santos, 2012; Call & Tomasello, 2008). Primate social learning, traditions and culture have also been extensively reviewed (Caldwell & Whiten, 2010; Hopper & Whiten, 2012; Whiten, 2012a), but not from the perspective of understanding other minds, which is what I shall pursue later in the chapter.

One of the recent reviews on primate mindreading was entitled “Does the chimpanzee have a theory of mind? 30 years later” (Call & Tomasello, 2008). This represents a substantial period of research—indeed somewhat longer than the period since explicit studies of children’s theory of mind began—and it offers an intriguing history of scientific discoveries, with marked fluctuations in the conclusions drawn about whether primates have anything like a theory of mind, and if so what forms it takes.

Following Premack & Woodruff’s (1978) initial (but quite heavily critiqued) experimental attacks on the question, there was a considerable lag before more work appeared. On the one hand this took the form of observations on spontaneous social behavior, such as what Richard Byrne and I (Whiten & Byrne, 1988a,b) called “tactical deception,” which revealed episodes in which monkeys and apes appeared to take into account what others could or could not see, might remember (thus what they might “know” or not), or intend. These cast a different, naturalistic

perspective on the possibility of primate mindreading and laid the foundations for experimental investigations that were later completed to test the interpretations we offered. However, their purely observational basis meant they were limited in terms of the conclusions about mindreading they could support—further experimental testing was essential to move the field forward.

Accordingly, new attempts at experimental investigations began in earnest. Initially, Povinelli, Nelson, & Boysen (1990) offered some evidence that chimpanzees could discriminate when human interactants could or could not see something, but a whole series of extensively controlled studies then followed that delivered a negative verdict on this possibility. Chimpanzees given the choice of begging for food from one familiar person versus another would choose someone facing them rather than facing away, but the discrimination broke down when finer distinctions were required, such as between a blindfold over the eyes versus over the mouth (Povinelli & Eddy, 1996). These were quite surprising results, at least to some familiar with chimpanzees, because the discriminations that chimpanzees failed to make sometimes concerned what appear to us humans as starkly different configurations: indeed some of the images of such contrasts, between a person with a bucket over their head versus the bucket being held, have become famous text book illustrations. Studies of monkeys that tackled similar questions came to a convergent, negative conclusion in relation to the possibility of anything like psychological attribution in primates (Cheney & Seyfarth, 1991)—after all, recognizing the distinction between seeing and not seeing something was recognized as one of the earliest stages in the development of children's understanding of mind (Flavell, Everett, Croft, & Flavell, 1981).

In their substantial 1997 review of primate cognition, Tomasello and Call accordingly delivered an essentially negative view on the likelihood of primate theory of mind, concluding that “there is no solid evidence that non-human primates understand the intentionality or states of mind of others” (p. 340). However, all changed over the next decade, in large part because of a new set of studies by Tomasello and Call's own research group. In a 2008 review, the same authors now concluded that “there is solid evidence from several different experimental paradigms that chimpanzees understand the goals and intentions of others, as well as the perception and knowledge of others” (Call & Tomasello, 2008, p. 187).

The picture began to change with a new approach that paired chimpanzees not with humans, but with other chimpanzees, and also put them into the kind of competitive situations that characterized the earlier tactical deception database. Here, chimpanzees were faced with a situation where they had a chance to grab one of two attractive food items before a more dominant chimpanzee was released head-on into the same enclosure a moment later (Hare, Call, Agnetta, & Tomasello, 2000). Although the subject could see both items, one of them was hidden from the dominant's perspective by a small opaque screen. The subject tended to choose the food item hidden from the dominant, but did not show a preference when the screen was made transparent. This suggests that, consistent with some of the observations of spontaneous deception that involved hiding some part of themselves from another individual, chimpanzees recognize the key geometric configurations that amount to what we normally distinguish as someone being able to see something, or not being able to see it (Bräuer, Call, & Tomasello, 2007; Whiten, 2013).

Later, Hare, Call, & Tomasello (2006) more directly tested the use of this ability in deception, confirming that chimpanzees would tend to steal a food item from a human on a side that provided them with cover from the person's visual field, rather than a transparent option on the other side. A similar discrimination was made in relation to stealing food where one option made a noise that would alert the person to what was going on, versus another that allowed a discrete and quiet approach (Hare et al., 2006). Chimpanzees thus appear to take into account the effects of their behavior not only on what other apes can or cannot see, but what they can or cannot hear.

A further development of the conspecific competition paradigm showed chimpanzees making adaptive distinctions when they had to remember which of two locations the dominant protagonist would have seen earlier, which amounts to taking into account what the protagonist **knew**, rather than simply what they could currently see (Hare, Call, & Tomasello, 2001). However, extensions of these and other paradigms to create tests of the attribution of false beliefs have drawn only negative results (Call & Tomasello, 1999; Kaminski, Call, & Tomasello, 2008), leading Call & Tomasello (2008) to note in their 30-year review that “despite several seemingly valid attempts, there is currently no evidence that chimpanzees understand false beliefs” (p. 187). However, taking into account these and also the more positive results of the studies reviewed above, they arrived at a significantly different stance than they had in 1997: “Our conclusion for the moment is, thus, that chimpanzees understand others in terms of a perception-goal psychology, as opposed to a full-fledged, human-like belief-desire psychology” (p. 187).

This statement draws on additional studies addressing not only the attribution of informational states like seeing, knowing and believing, but also the attribution of intentions. An important study concerning the attribution of intentions drew on the spirit of Whiten’s (1996) proposal that one powerful way to identify the attribution of states of mind in a non-verbal creature would be to assume the mindreader recognizes these as intervening variables operating in the mindread individual, computed on the basis of multiple, alternative, observable circumstances faced by that individual, that help predict (and/or explain) its responses in multiple, alternative contexts. Call, Hare, Carpenter, & Tomasello (2004) engineered multiple contexts in which a chimpanzee might infer an interactant was either unwilling or unable to complete an offer of food they had begun, and obtained positive results that thus implied a capacity to attribute intent. This represents one important component in the “perception-goal” psychology alluded to above.

Cultural transmission—from mind to mind

Mindreading makes for a society of interconnected minds in an obvious and explicit way: indeed, we might say that true mindreading is defined by such connections. Because humans become such out-and-out mentalists during their childhoods, doing such things as using language replete with mental state terminology to discuss their own and others’ minds, it seems entirely justified to describe human groups as constituted by interconnected minds, a core aspect of deep social mind.

Can we really say something similar of culture and the psychological attributes that make it possible? One important affirmative answer is that culture exemplifies deep social mind insofar as the contents of our minds include vast swathes of information inherited from other minds, in some cases from minds long dead, but passed on to us through a sometimes extremely long series of intervening generations—as illustrated by such miscellaneous examples as the concepts of the spear, the wheel, and gods. Our minds are clearly deeply socially constituted, in this respect, but does this process imply understanding other minds? I suggest the answer must be: “not necessarily.” Traditions are now known to pass from generation to generation in an expanding range of animals, including primates and other mammals, birds and fish (e.g. Thornton & Clutton-Brock, 2011; Whiten, Hinde, Stringer, & Laland, 2011, 2012). For example, travel routes have been shown by translocation experiments to be inherited by social learning in several species of fish (Laland, Atton, & Webster, 2011). In such a process, some form of representation of the route concerned must thus pass with sufficient fidelity from the brains of the first set of fishes to those of the fish who acquire the routes from them, and to this extent we have an “inter-mental” process—information has somehow leapt from brain to brain. However, the fish who inherit the routes do not

need to understand or “read” the minds of their cultural ancestors to achieve this: they need only observe them or perhaps even only follow or join their shoal, to acquire the tradition.

The connectedness between minds that underlies cultural transmission may become more direct, however, when we begin to discriminate among forms of social transmission and focus on the most sophisticated. Two in particular are relevant—intentional teaching and imitation. Intentional teaching achieves this status insofar as it relies minimally on an ability to recognize what a potential pupil does or does not know, and thus requires teaching about. More complex attributions could include what the pupil believes, or even falsely believes.

The link between imitation and explicit mindreading is more subtle and complex and addressed in some rather different ways by different authors. A fundamental point is that mindreading and imitation require a translation between the perspectives of others and oneself—in both, one needs in some sense to be able to “stand in the shoes” of the other to make the appropriate translation between other and self (Whiten, 1996). In mindreading, one needs to re-represent in one’s brain the mental state of another, such as their false beliefs. In imitation, one needs to translate from one’s perception of what it is for another individual to, for example, tie their shoelace in a certain way, to what it will be for you to generate a behavioural copy of this, from your own quite different perspective. The analogy is perhaps particularly pertinent in the case of “simulation” theories of mindreading, in which one succeeds by putting oneself in the position of the other and in effect imagining what it is like to “be” them (Stone & Davies, 1996). Empathy is another case where the link to imitation is evident, for in empathy, in a sense, one imitates the emotional state of another individual (Iacoboni, 2012). In line with such considerations, Meltzoff & Gopnik, 1993; Meltzoff, 2005) proposed that imitation in infancy provides a developmental precursor to theory of mind, because it allows the infant to associate states of mind (its own) with the corresponding behavioural manifestations seen in others. From the different perspective of autism research, Rogers & Pennington (1991) suggested that early deficits in imitation represent the start of a cascade of problems in autism that are later manifested in delays in theory of mind achievements. Evidence for an early deficit in imitation in autism is extensive (Williams, Whiten, & Singh, 2004) and there is also evidence consistent with the proposed linkage with theory of mind difficulties (Perra, Williams, Whiten, Fraser, Benzie, & Perrett, 2008), although this remains a controversial and contested area of research (Rogers & Williams, 2006; Southgate & Hamilton, 2008).

Research on mirror neurons has suggested a specific neural substrate relevant to these theories, for mirror neurons have been implicated in imitation in humans (Iacoboni, Woods, Brass, Bekkering, Mazziotta, & Rizzolatti, 1999; Iacoboni, 2012), as well as recognition of others’ goals in monkeys (Umla, Kohler, Gallese, Fogassi, Fadiga, Keysers, et al., 2001; Rizzolatti, 2005), leading to hypotheses concerning a role for mirror neurons in the evolution of mindreading (Gallese & Goldman, 1998; Gallese, 2005, 2005). Williams, Whiten, Suddendorf, & Perrett (2001) brought several of the above issues together to suggest links between distorted mirror neuron function and both imitation and theory of mind deficits in autism (see Dapretto, Davies, Pfeifer, Scott, Sigman, Bookheimer, et al., 2006).

The above considerations suggest that attempts to reconstruct and understand the evolution of “understanding other minds” should, in the case of cultural transmission, focus particularly on the processes of imitation and teaching. Further below I do this, including examining evidence that any such processes suggest selectivity in relation to such psychological elements as the intentions and rationality of different actions in others. More broadly, aspects of cultural transmission relevant to our focus on “Deep Social Mind” include such phenomena as contagion and conformity (doing what a majority of others are doing, just because others are acting this way: Claidiere & Whiten, 2012b) and over-imitation (faithfully copying others despite blatant evidence the acts are

not effective: Lyons, Young, & Keil, 2007; Lyons, Damrosch, Lin, Macris, & Keil, et al., 2011; Whiten et al., 2009a), that imply a special power of elements of the social world to be overridingly influential and deeply penetrate an individual's decision-making.

Reconstructing ancestral cultural capacities

The most recent phases of the evolution of human cultures and cultural capacities can be traced through a variety of complementary methods including archaeology, cultural phylogenetics and studies of the core characteristics of the hunting and gathering way of life identified across multiple instances of such societies today and in the recent past (Whiten et al., 2011, 2012). To trace earlier evolutionary phases in the common ancestor we share with other primates, a commonly accepted procedure is to test for elements of culture shared across a clade or family of related species such as the apes, and accordingly attribute these elements to inheritance from the appropriate common ancestor of that group (Byrne, 1995).

In doing this for the last common ancestor we shared with chimpanzees approximately 6 million years ago, or with the ancestor shared with all the great apes around 14 million years ago, I have considered in turn three main aspects of culture: (i) the range of traditions and their distribution in time and space; (ii) the behavioural contents of the traditions; and (iii) the underlying transmission mechanisms (Whiten, 2005, 2011). Here, the topic of this volume leads me to focus mostly on the latter, first treating the other two relatively briefly; for a fuller treatment see Whiten (2011).

The range of ape traditions and their distributions have been inferred on the basis of collated reports across multiple long-term field studies for two species in particular. For chimpanzees, these delineated 39 different traditions spanning food processing, tool use and forms of social and sexual behavior (Whiten, Goodall, McGrew, Nishida, Reynolds, Sugiyama, et al., 1999; Whiten, Goodall, McGrew, Nishida, Reynolds, Sugiyama, et al., 2001; Whiten, 2005), a corpus that has expanded with further studies (McGrew, 2004; Whiten, 2010). Chimpanzees can be assigned to a particular locality on the basis of the unique cultural profile they display from among these possibilities, as can humans. A similar picture obtains among orangutans, if perhaps a little less rich in its manifestations (van Schaik, Ancrenaz, Borgen, Galdikas, Knott, Singleton, et al., 2003; van Schaik & Burkart, 2011), so that the existence of a cultural life constituted by multiple traditions of varied behavioural kinds can likely be attributed not only to our last common ancestor with chimpanzees, but to the great ape ancestral stock of around 14 million years ago. The inferences drawn from observational studies of wild apes have been supported by experiments with captive apes demonstrating a capacity to sustain traditions through repeated transmission events, extending across several communities in the case of chimpanzees (Whiten, Horner, & de Waal, 2005; Whiten, Spiteri, Horner, Bonnie, Lambeth, Schapiro, et al., 2007; Horner, Whiten, Flynn, & de Waal 2006; for comparable child studies see Whiten & Flynn, 2010; Flynn & Whiten, 2012).

The behavioural *contents* that make up these ape cultures span most of the principal domains of ape behavior. In the case of chimpanzees they include a remarkable range of kinds of tool use that is shared with (although of course vastly exceeded by) our own species and these include some particular forms of special interest to students of human evolution, notably hammer and anvil use (to crack nuts) and other forms of percussive and force-based tool use (Whiten et al., 2009b) that indicate important foundations for the tool knapping and other stone age cultures that occupied most of the last two million years of *Homo* history.

Imitation and other cultural transmission processes.

"Do apes ape?" asked Tomasello (1996). The answer to this question depends as much upon how imitation is defined as on the relevant empirical data. Whiten & Ham (1992) defined imitation as

copying the form of another individual's action. This was later contrasted with "emulation," in which instead of copying the actions of the other individual, an observer learns from the results of their actions and then seeks to recreate these (Tomasello, 1990, 1996). Some researchers distinguish imitation from emulation by insisting that imitation must involve bodily copying. Others instead allow that the "form" of another's actions that an imitator might copy include a particular sequential or hierarchically-organized program of action subcomponents (Byrne & Russon, 1998), which might be describable in terms of bodily actions, but might also involve tool use or other objects moved in particular ways.

Imitation defined as bodily copying is often hard to distinguish from emulation because an ape action that is going to serve as an experimental model typically gains rewards by acting on the world, and the results of this are thus likely to be confounded with the bodily movements responsible. One way this has been circumvented is to train a participant to act on command ("do this") in a "Simon says" game, in which after the participant has grasped the copying ground rule through a series of training actions, it is tested on a battery of relatively novel actions. Both chimpanzees and an orangutan have been shown capable of mastering this game and generating a significant number of matches, recognizable by coders blind to what model action the participant has just seen (Custance, Whiten, & Bard, 1995; Call, 2001). The battery used in both studies required a range of bodily imitations including facial, gestural and whole body instances. Interestingly, only these apes have successfully mastered the essential copying idea behind the game played, whereas several studies of this kind have failed to achieve this with monkeys. This implies that only the apes are capable of understanding the underlying concept of imitation—they can tell when they are imitating or not—expressing a distinctive level of attunement with the model (in these studies, another species of ape—a human—with a body morph somewhat similar yet in some ways different to that of the participant).

In a study testing social learning of nut-cracking behavior in chimpanzees naïve to the technique, Marshall-Pescini & Whiten (2008) recorded several sequences in which the young observer participants, while holding no hammer stone nor having any nut to crack, began spontaneously to produce hammering actions with their arms whilst watching the already-proficient model hammering a nut open. In some cases a rough behavioural synchrony of observer and model was observed (for a video clip, see the electronic supplementary information to Marshall-Pescini & Whiten, 2008). This may suggest that the observed actions were being coded in the observer chimpanzee in action terms, consistent with the operation of mirror neurons. This is an interesting possibility, because on the one hand, we know of mirror neurons through studies of monkeys, rather than apes (because of ethical reasons for not pursuing invasive single-cell recording with apes), and these neurons have been argued not to be involved in imitation, because the monkeys concerned do not show evidence of imitation; instead they have been suggested to function in the interpretation of other's goal-direction actions (Rizzolatti, 2005). In apes, mirror neurons may have been co-opted into their imitative capacities, as appears to be the case for humans (Iacoboni, 2012).

Other reasons to discount any simple dichotomy of the kind "apes emulate, humans imitate" include experiments explicitly testing for emulation learning through the use of "ghost conditions" in which there is no model to copy, the movements of objects that would normally be manipulated by a model being instead engineered through artificial means and thus displaying only the environmental results of (hidden) actions that an emulator is supposed to learn from (Hopper, 2010). Hopper, Spiteri, Lambeth, Schapiro, Horner, & Whiten (2007) found that when the task was a complex and challenging tool use one, chimpanzees could not learn from such a ghost condition. In this task, food was stuck behind a blockage, which could be lifted up using a stick tool to

release the food, and in the ghost condition this was made to happen through the block and/or tool being made to move by the pulling of hidden fishing line, with no model chimpanzee involved. By contrast, when they saw a chimpanzee use the tool to free the blockage they would copy this, with sufficient fidelity to generate different tool use traditions when the initial models applied different techniques (Whiten et al., 2005). This suggests that in such situations they need to see what is done with a tool, by an agent. However, with a simpler, manual task in which a small door was simply slid to one side or the other to reveal a food reward, chimpanzees did learn in a ghost condition (Hopper, Lambeth, Schapiro, & Whiten, 2008). Together with other recent evidence of emulation (Tennie, Call, & Tomasello, 2010), this suggests that these apes have a portfolio of social learning skills that are deployed differentially according to context. Further, direct evidence of this comes from a study by Horner & Whiten (2005) in which young chimpanzees showed a tendency to imitate the main parts of a sequence of actions to extract a food reward from a box when this was opaque and they could not see that certain actions were actually causally irrelevant, but missed these out when a transparent version of the box was involved, thus taking a more emulative approach.

This selectivity in relation to contextual information about physical causality has also been found in relation to information about what we might call psychological causality, linking again to the topic of MR. Here, Buttelmann, Carpenter, Call, & Tomasello (2007) designed tests modeled on those of Gergely, Bekkering, & Kiraly (2002) who had shown that human infants, who would readily imitate an adult using the bizarre action of butting a light with their head to switch it on so long as this person's hands were free, were rarely willing to do so if the hands were occupied. This implied that infants operate a quite sophisticated model of human action that recognizes the conditions under which actions may be freely chosen, intentional, and thus worth copying even if superficially rather odd. Buttelmann and colleagues showed that chimpanzees also made this distinction, in relation to three different scenarios in which human models used their head, foot or bottom to switch on a light or sound (sitting on an object, in the latter case), either with their hands occupied (so using other means was the rational thing to do) or hands free (so using other means indicated this was an intentionally chosen technique). Tomasello & Carpenter (2005) also showed that like human infants, when presented with human models attempting, but failing to achieve a certain outcome, chimpanzees would themselves then attain this outcome, apparently recognizing and completing the model's perceived intentions. This evidence converges with that of Call et al. (2004) reviewed above in the section on evidence of MR in apes.

Another aspect of social learning that underlines the power of the social world in shaping behavior is conformity. Conformity is doing what others do just because others are doing it, and in particular, following the majority. In humans, this tendency was graphically illustrated long ago by social psychology experiments in which participants were asked to make perceptual judgments, such as which of three lines was longest, in a group context (Asch, 1956). Unbeknownst to the participant, other members of the group were confederates of the experimenter, and in some conditions of the experiment they all consistently judged one line longer that was patently not: nevertheless a significant number of participants followed suit, demonstrating strong conformity. Recent studies reviewed by Claidiere & Whiten (2012) have shown this tendency is not restricted to humans, but has occurred in our own studies of primates, including chimpanzees (Whiten et al., 2005; Dindo, de Waal, & Whiten, 2009) and in a variety of other species including fish (Pike & Laland, 2010).

I have recently suggested that another aspect of social learning—"over-imitation"—shares with conformity the central feature of allowing social information to override personal information when the two are in conflict (Whiten (2012b)). In over-imitation, a child copies aspects of another

individual's behavior even though these can be readily seen to have no causal significance for the outcome at stake. This was evident when Horner & Whiten (2005) repeated the experiment described above in which a model performed both causally relevant and causally irrelevant actions on either an opaque box or a transparent box. Whilst chimpanzees have behaved differently on witnessing either the transparent or the opaque condition, imitating only in the latter case, children were found to copy in blanket fashion in both conditions. Lyons et al. (2007, 2011) have replicated these results with larger sample sizes and additional controls, and dubbed the phenomenon "over-imitation." Tested in this way, it has appeared to be a characteristic of human social learning, extending to adults (McGuigan, Makinson, & Whiten, 2011), that we did not see in chimpanzees. However, one recent study did suggest an allied phenomenon. In this study, chimpanzees were shown to learn by observation to construct a long stick-tool from two smaller ones in order to gain some out-of-reach food (Price, Lambeth, Schapiro, & Whiten, 2009). However, a handful of control participants, who had seen no model, managed to work this out for themselves. The intriguing result emerged when the chimpanzees were later tested with the food closer so that the long tool was not needed: now, the individual learners gave up using the long tool; but the social learners persevered in using it (awkwardly), demonstrating yet again the power of socially-gained information to override the personal information available.

Conclusions: evolutionary foundations of interconnected minds

Before one can discuss evolutionary foundations, one has to have a good account of what it is that one is investigating the foundations of. In the case of this volume, the short answer is of course the topic of "understanding other minds," typically investigated over the last 30 years or so as "theory of mind," or "mindreading." However, a richer and broader literature has come to suggest that MR is but one part of a larger complex of forms of social cognition that I outlined in the first part of this chapter, that together constitute what I called "Deep Social Mind", incorporating MR, culture, cooperation and language. Once this adaptive complex is sufficiently well specified, one can begin to investigate evolutionary origins, using the comparative method to identify fundamental phenomena shared by ourselves and our closest primate relatives—the subject of the remainder of the chapter.

Elsewhere Erdal and I offer an evolutionary analysis ranging over all four of the principal "pillars" of Deep Social Mind outlined above (Whiten and Erdal, 2012), but here I have focused on the two that have the most particular relevance to the present volume; first MR, and second, the socio-cognitive underpinnings of culture. The latter topic has occupied the greater part of our most recent research efforts and accordingly has been discussed in the most depth here.

The last dozen or so years of research in primate MR have been an exciting period witnessing something of a sea-change in what new experimental evidence has taught us. At the turn of the century such evidence was largely negative, but the spate of new investigations, highlights of which were outlined above, have "converted" two of the original skeptics to conclude that apes are best characterized as "perception-goal" psychologists (Call & Tomasello, 2008). This conclusion comes full circle to achieve consistency with some of the inferences made on the basis of much earlier observational studies of wild and captive primates (see also Suddendorf & Whiten, 2001 and (Whiten, 2013)). Humphrey (1980) for example, suggested some primates were "natural psychologists"; and now the experimental work is circumscribing the scope of that natural psychology, concerned with understanding some basic aspects of perception, or seeing, and goals and intentions. In the first edition of *Understanding Other Minds*, reviewing our studies of primate deception and counter deception, I tentatively concluded that "some chimpanzees at least seem

capable of discriminating between the apparent and real *intentions* of others that occur in cases of potential deception” (Whiten, 1993, p 375; Whiten & Byrne, 1988a,b), as well as discriminating when another individual can see something or not, or is likely to notice something or not (Whiten, 1993, p 377–8).

There is, thus, something of a consensus spanning field observations and experiments with captive individuals, that apes achieve a significant attunement with the minds of others, in circumstances where the others’ psychological states—their view of the world, their motivations and intentions—may be quite different to the self’s (Whiten, 2013). Of course, this is not to say that apes conceive of other minds as such: their achievements in this sphere can equally be seen as a sophisticated kind of behaviour-reading—but then, humans are not telepathists either and must read psychological states like seeing and wanting through observables (in the case of seeing, based on observable geometry, transparency, eyes open and so on). These are the means through which attunement to other minds occurs.

It follows that it will be this kind of self-other attunement that is available to apes when we shift our focus onto imitation and kindred kinds of observational learning from others. Here, attunement likewise has to bridge between other’s and self’s perceptual and motoric engagement with the world. The motivation to “be like you” can be very strong, as indicated by such evidence of conformity as was reviewed above—although to be sure, this motivation becomes yet stronger in our own species, as we see in the phenomenon of over-imitation. At the same time, we have seen that apes do not “mindlessly ape” others: copying is selective in several ways, discriminating such variables as physical causality in what a model is manipulating, the intentions of the model, and the constraints under which they operate. Consistent with the account offered in this chapter, a recent book on *The Primate Mind* (de Waal & Ferrari, 2012) was subtitled *Built to connect with other minds*. The evidence reviewed here suggests that in living apes, and by inference in our common ape ancestors, the connections can be both strongly motivated and sophisticated in their discriminative architecture. This would have provided significant foundations for the elaborate understanding of minds we human mentalists achieve.

References

- Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychology Monographs* 70: 1–70.
- Avis, J. & Harris, P. L. (1991). Belief-desire reasoning among Baka children: evidence for a universal conception of mind. *Child Development* 62: 460–7.
- Bräuer, J., Call, J., & Tomasello, M. (2007). Chimpanzees really know what others can see in a competitive situation. *Animal Cognition* 10: 439–48.
- Buttelmann, D., Carpenter, M., Call, J., & Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental Science* 10: 31–8.
- Byrne, R. W. (1995). *The Thinking Ape: Evolutionary Origins of Intelligence*. New York: Oxford University Press.
- Byrne, R. W. & Russon, A. E. (1998). Learning by imitation: a hierarchical approach. *Behavioral and Brain Sciences* 21: 667–721.
- Caldwell, C. A. & Whiten, A. (2010). Social learning in monkeys and apes: cultural animals? In: C. Campbell, A. Fuentes, K. MacKinnon, S. Bearder, & R. Stumpf, (Eds), *Primates in Perspective* 2nd edn, pp. 652–62. Oxford: Oxford University Press.
- Call, J. (2001). Body imitation in an enculturated orangutan (*Pongo pygmaeus*). *Cybernetics and Systems: An International Journal* 32: 97–119.

- Call, J., Hare, B. H., Carpenter, M., & Tomasello, M. (2004). "Unwilling" versus "unable": Chimpanzees' understanding of human intentional action? *Developmental Science* 7: 488–98.
- Call, J. & Santos, L. R. (2012). Understanding other minds. In: J. Mitani, J. Call, P. Kappeler, R. Palombit, & J. Silk (Eds), *The Evolution of Primate Societies* (pp. 664–81). Chicago: Chicago University Press.
- Call, J. & Tomasello, M. (1999). A non-verbal false belief task: the performance of children and great apes. *Child Development* 70: 381–95.
- Call, J. & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Science* 12: 187–92.
- Cheney, D. R. and Seyfarth R. M. (1991). Reading minds or reading behavior? Tests for a theory of mind in monkeys. In: A. Whiten (Ed.), *Natural Theories of Mind; Evolution, Development and Simulation of Everyday Mindreading* (pp. 175–94). Oxford: Basil Blackwell.
- Claidière, N. & Whiten, A. (2011). Integrating the study of conformity and culture in humans and non-human animals. *Psychological Bulletin* 138(1): 126–45.
- Custance, D. M., Whiten, A., & Bard, K. A. (1995). Can young chimpanzees (*Pan troglodytes*) imitate arbitrary actions? Hayes and Hayes (1952) revisited. *Behaviour* 132: 837–59.
- Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., & Iacoboni, M. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience* 9: 28–30.
- De Waal, F. B. M. & Ferrari, P. F. (Eds) (2012). *The Primate Mind: Made to Connect with Other Minds*. Cambridge: Harvard University Press.
- Dindo, M., de Waal, F. B. M., & Whiten, A. (2009). In-group conformity sustains different foraging traditions in capuchin monkeys (*Cebus apella*). *PLoS One* 4: e7858. Available at: <http://dx.plos.org/10.1371/journal.pone.0007858>
- Erdal, D., & Whiten, A. (1994). On human egalitarianism: an evolutionary product of Machiavellian status escalation? *Current Anthropology* 35: 175–83.
- Erdal, D., & Whiten, A. (1996). Egalitarianism and Machiavellian intelligence in human evolution. In: P. Mellars and K. Gibson (Eds), *Modelling the Early Human Mind* (pp. 139–50). Cambridge: McDonald Institute Monographs.
- Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: further evidence for the level 1–level 2 distinction. *Developmental Psychology* 17: 99–103.
- Flynn, E. G., & Whiten, A. (2012). Experimental "microcultures" in young children: identifying biographic, cognitive and social predictors of information transmission. *Child Development* 83: 911–25.
- Gallese, V. (2005). "Being like me": Self-other identity, mirror neurons and empathy. In: S. Hurley and N. Chater (Eds) *Perspectives on Imitation, Volume 1, Mechanisms of Imitation and Imitation in Animals* (pp. 101–18). Cambridge: MIT Press.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Science* 2: 493–501.
- Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature* 415: 755.
- Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour* 59: 771–86.
- Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour* 61: 139–51.
- Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition* 101: 495–514.
- Hewlett, B. S., Fouts, H. N., Boyette, A., & Hewlett, B. L. (2011). Social learning among Congo Basin hunter-gatherers. *Philosophical Transactions of the Royal Society B* 366: 1168–78.
- Hopper, L. M. (2010). "Ghost" experiments and the dissection of social learning in humans and animals. *Biological Reviews* 85: 685–701.

- Hopper, L. M., Lambeth, S. P., Schapiro, S. J., & Whiten, A. (2008). Observational learning in chimpanzees and children studied through “ghost” conditions. *Proceedings of the Royal Society, London B* 275: 835–40.
- Hopper, L. M., Spiteri, A., Lambeth, S. P., Schapiro, S. J., Horner, V., & Whiten, A. (2007). Experimental studies of traditions and underlying transmission processes in chimpanzees. *Animal Behaviour* 73: 1021–32.
- Hopper, L. M. & Whiten, A. (2012). The comparative and evolutionary psychology of social learning and culture. In: J. Vonk & T. Shackelford (Eds), *The Oxford Handbook of Comparative Evolutionary Psychology* (pp. 451–73). Oxford University Press.
- Horner, V. K. & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children. *Animal Cognition* 8: 164–81.
- Horner, V., Whiten, A., Flynn, E., & de Waal, F. B. M. (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences, USA* 103: 13878–83.
- Humphrey, N. K. (1980). Nature’s psychologists. In: B. Josephson & V. Ramachandran (Eds), *Consciousness and the Physical World* (pp. 57–80). London: Pergamon Press.
- Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science* 286: 2526–8.
- Iacoboni, M. (2012). The human mirror system and its role in imitation and empathy. In: F. B. M. de Waal & P. F. Ferrari (Eds), *The Primate Mind: Built to Connect with Other Minds* (pp. 32–47). Cambridge: Harvard University Press.
- Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition* 109: 224–34.
- Laland, K. N., Atton, N. & Webster, M. M. (2011). From fish to fashion: experimental and theoretical insights into the evolution of culture. *Philosophical Transactions of the Royal Society B* 366: 958–68.
- Lee, R. B. & DeVore, I. (Eds) (1968). *Man the Hunter*. Chicago: Aldine de Gruyter.
- Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin* 123: 3–32.
- Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences, USA* 104: 19751–6.
- Lyons, D. E., Damrosch, D. H., Lin, J. K., Macris, D. M., & Keil, F. C. (2011). The scope and limits of overimitation in the transmission of artefact culture. *Philosophical Transactions of the Royal Society B* 366: 1158–67.
- Marlowe, F. W. (2005). Hunter-gatherers and human evolution. *Evolutionary Anthropology* 14: 54–67.
- Marshall-Pescini, S., & Whiten, A. (2008). Social learning of nut-cracking behavior in East African sanctuary-living chimpanzees (*Pan troglodytes schweinfurthii*). *Journal of Comparative Psychology* 122: 186–94.
- McGrew, W. C. (2004) *The Cultured Chimpanzee: Reflections on Cultural Primatology*. Cambridge: Cambridge University Press.
- McGuigan, N., Makinson, J., & Whiten, A. (2011). From over-imitation to super-copying: Adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. *British Journal of Psychology* 102: 1–18.
- Meltzoff, A. N. (2005). Imitation and other minds: the “like me” hypothesis. In: S. Hurley & N. Chater (Eds), *Perspectives on Imitation, Volume 2, Imitation, Human Development and Culture* (pp. 55–77). Cambridge: MIT Press.
- Meltzoff, A. N. & Gopnik, A. (1993). The role of imitation in understanding persons and developing a theory of mind. In: S. Baron-Cohen, H. Tager-Flusberg, & J. D. Cohen (Eds), *Understanding Other Minds: Perspectives from Autism* (pp. 335–66). Oxford: Oxford University Press.
- Perra, O., Williams, J. H. G., Whiten, A., Fraser, L., Benzie, H. & Perrett, D. I. (2008) Imitation and “theory of mind” competencies in discrimination of autism from other neurodevelopmental disorders. *Research in Autism Spectrum Disorders* 2: 456–68.

- Pike, T. W., & Laland, K. N. (2010). Conformist learning in nine-spined sticklebacks' foraging decisions. *Biology Letters* 6(4): 466–8.
- Povinelli, D. J. & Eddy, T. J. (1996) What young chimpanzees know about seeing. *Monographs of the Society Research in Child Development* 61: 1–152.
- Povinelli, D. J., Nelson, K. E., & Boysen, S. T. (1990). Inferences about guessing and knowing by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology* 104: 203–10.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1: 515–26.
- Price, E. E., Lambeth, S. P., Schapiro, S. J., & Whiten, A. (2009). A potent effect of observational learning on chimpanzee tool construction. *Proceedings of the Royal Society B* 276: 3377–83.
- Rizzolatti, G. (2005). The mirror neuron system and imitation. In: S. Hurley & N. Chater (Eds), *Perspectives on Imitation, Volume 1, Mechanisms of Imitation and Imitation in Animals* (pp. 55–76). Cambridge: MIT Press.
- Rogers, S. J., & Pennington, B. F. (1991). A theoretical approach to the deficits in infantile autism. *Developmental Psychopathology* 3: 137–62.
- Rogers, S. J. & Williams, J. H. G. (2006). *Imitation and the Social Mind: Autism and Typical Development*. New York: Guilford Press.
- Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology* 47: 1239–47.
- Southgate, V. & Hamilton, A. F. de C. (2008). Unbroken mirrors: challenging a theory of autism. *Trends in Cognitive Science* 12: 225–9.
- Stone, T., & Davies, M. (1996). The mental simulation debate: a progress report. In: P. Carruthers & P. K. Smith (Eds) *Theories of Theories of Mind* (pp. 119–37). Cambridge: Cambridge University Press.
- Suddendorf, T., & Whiten, A. (2001). Mental evolution and development: Evidence for secondary representation in children, great apes and other animals. *Psychological Bulletin* 127: 629–50.
- Tennie, C., Call, J., & Tomasello, M. (2010) Evidence for emulation in chimpanzees in social settings using the floating peanut task. *PLoS ONE* 5, e10544.
- Thieme, H. (1997). Lower Paleolithic hunting spears from Germany. *Nature* 385: 807–10.
- Thornton, A., & Clutton-Brock, T. (2011). Social learning and the development of individual and group behavior in mammal societies. *Philosophical Transactions of the Royal Society B* 366: 978–87.
- Thornton, A., & Clutton-Brock, T. (2011). Social learning and the development of individual and group behavior in mammal societies. *Philosophical Transactions of the Royal Society B* 366: 978–87.
- Tomasello, M. (1990). Cultural transmission in the tool use and communicatory signalling of chimpanzees? In: S. Parker & K. Gibson (Eds), *Language and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives* (pp. 274–311). Cambridge: Cambridge University Press.
- Tomasello, M. (1996). Do apes ape? In: C. M. Heyes & B. G. Galef (Eds), *Social Learning in Animals: The Roots of Culture* (pp. 319–46). London: Academic Press.
- Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge: Harvard University Press.
- Tomasello, M., & Call, J. (1997). *Primate Cognition*. Oxford: Oxford University Press.
- Tomasello, M., & Carpenter, M. (2005). The emergence of social cognition in three young chimpanzees. *Monographs of the Society for Research in Child Development* 70 (1, Serial No. 279).
- Tooby, J., & DeVore, I. (1987) The reconstruction of hominid behavioral evolution through strategic modelling. In: W. G. Kinzey (Ed.), *The Evolution of Human Behavior: Primate Models* (pp. 183–227). New York: SUNY Press.
- Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I know what you are doing: a neuropsychological study. *Neuron* 31: 155–65.
- van Schaik, C. P. (2009). Geographic variation in the behavior of wild great apes: is it really cultural? In: K. N. Laland & B. G. Galef (Eds), *The Question of Animal Culture* (pp 70–98). Cambridge: Harvard University Press.

- van Schaik, C. P., & Burkart, J. M. (2011). Social learning and evolution: The cultural intelligence hypothesis. *Philosophical Transactions of the Royal Society B* 367: 1008–16.
- van Schaik, C. P., Ancrenaz, M., Borgen, G., Galdikas, B., Knott, C. D., Singleton, I., Suzuki, A., Utami, S. S., & Merrill, M. (2003). Orangutan cultures and the evolution of material culture. *Science* 299, 102–5.
- Wellman, H. M., Fang F. X., & Peterson, C. C. (2011). Sequential progressions in a theory of mind scale: longitudinal perspectives. *Child Development* 82: 780–92.
- Whiten A. (1993). Evolving a theory of mind: The nature of non-verbal mentalism in other primates. In: S. Baron-Cohen, H. Tager-Flusberg, D. Cohen, & F. Volkmar (Eds), *Understanding Other Minds: Perspectives from Autism* (pp. 367–96). Oxford: Oxford University Press.
- Whiten, A. (1996). When does behavior reading become mindreading? In P. Carruthers and P. K. Smith (Eds), *Theories of Theories of Mind* (pp. 277–92). Cambridge: Cambridge University Press.
- Whiten, A. (1999). The evolution of deep social mind in humans. In: M. Corballis & S. E. G. Lea (Eds), *The Descent of Mind* (pp. 155–75). Oxford: Oxford University Press.
- Whiten, A. (2005). The second inheritance system of chimpanzees and humans. *Nature* 437: 52–5.
- Whiten, A. (2006). The place of “deep social mind” in the evolution of human nature. In M. A. Jeeves (Ed.), *Human Nature* (pp. 207–22). Edinburgh: Royal Society of Edinburgh.
- Whiten, A. (2010). A coming of age for cultural Panthropology. In: E. Lonsdorf, S. Ross & T. Matsuzawa (Eds), *The Mind of the Chimpanzee*. Chicago: Chicago University Press.
- Whiten, A. (2011). The scope of culture in chimpanzees, humans and ancestral apes. *Philosophical Transactions of the Royal Society B* 366: 997–1007.
- Whiten, A. (2012a). Primate social learning, traditions and culture. In: J. Mitani, J. Call, P. Kappeler, R. Palombit & J. Silk (Eds), *The Evolution of Primate Societies* (pp. 682–700). Chicago: Chicago University Press.
- Whiten, A. (2012b). Social cognition: making us smart, or sometimes making us dumb? Overimitation, conformity, non-conformity and the transmission of culture in ape and child. In: M. Banaji & S. Gelman (Eds), *The Development of Social Cognition* (pp. 150–4). Oxford: Oxford University Press.
- Whiten, A. (2013). Humans are not alone in computing how others see the world. *Animal Behaviour* 86 (in press).
- Whiten, A., & Byrne, R. W. (1988a). Tactical deception in primates. *Behavioral and Brain Sciences* 11: 233–73.
- Whiten, A., & Byrne, R. W. (1988b). The manipulation of attention in primate tactical deception. In: R. W. Byrne & A. Whiten (Eds), *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans* (pp. 211–23). Oxford: Oxford University Press.
- Whiten, A., & Erdal, D. (2012). The human socio-cognitive niche and its evolutionary origins. *Philosophical Transactions of the Royal Society B* 367: 2119–29.
- Whiten, A., & Flynn, E. G. (2010). The transmission and evolution of experimental “microcultures” in groups of young children. *Developmental Psychology* 46: 1694–709.
- Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., Tutin, C. E. G., Wrangham, R. W., & Boesch, C. (1999). Cultures in chimpanzees. *Nature* 399: 682–5.
- Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., et al. (2001). Charting cultural variation in chimpanzees. *Behaviour* 138: 1481–516.
- Whiten, A., & Ham, R. (1992). On the nature and evolution of imitation in the animal kingdom: Reappraisal of a century of research. *Advances in the Study of Behaviour* 11: 239–83.
- Whiten, A., Hinde, R. A., Stringer, C. B., & Laland, K. N. (2011). Introduction. Culture eEvolves. *Philosophical Transactions of the Royal Society B* 366: 938–48.
- Whiten, A., Hinde, R. A., Stringer, C. B. & Laland, K. N. (Eds) (2012). *Culture Evolves*. Oxford: Oxford University Press.
- Whiten, A., Horner, V., & de Waal, F. B. M. (2005). Conformity to cultural norms of tool use in chimpanzees. *Nature* 437: 737–40.

- Whiten, A., McGuigan, H., Hopper, L. M. and Marshall-Pescini, S. (2009a). Imitation, over-imitation, emulation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B* 364: 2417–28.
- Whiten, A., Schick, K., & Toth, N. (2009b). The evolution and cultural transmission of percussive technology: integrating evidence from paleoanthropology and primatology. *Journal of Human Evolution* 57: 420–35.
- Williams, J. H. G., Whiten, A., and Singh, T. (2004). A systematic review of action imitation in autistic spectrum disorder. *Journal of Autism Developmental Disorders* 34: 285–99.
- Williams, J. H. G., Whiten, A., Suddendorf, T., & Perrett, D. I. (2001). Imitation, mirror neurons and autism. *Neuroscience and Biobehavioural Reviews* 25: 287–95.

Mindreading by simulation: The roles of imagination and mirroring

Alvin I. Goldman and Lucy C. Jordan

Criteria of adequacy for a theory of “theory of mind”

There is consensus in cognitive science that ordinary people are robust mindreaders and that mindreading begins early in life. Many other questions concerning mindreading, however, remain in dispute, including the four that follow:

- (1) By what method(s) do cognizers read other people’s minds—that is, attribute mental states to them? Which cognitive capacities, mechanisms, or processes play pivotal roles in mindreading?¹ Call this the task-execution question.
- (2) How did the human species, and how do individuals, acquire mindreading capacities? What are the phylogenetic and ontogenetic parts of the story? Call this the acquisition question.
- (3) What neural substrates underlie mindreading? In other words, does the proposed story pass neuroscientific muster? Call this the neural plausibility question.
- (4) How does the proffered story of mindreading mesh with the general story of human cognition? Is mindreading a typical example of cognition, or is it a singularity—a one-off piece of cognitive hardware? Call this the question of mesh.

Any theory or approach to mindreading must answer these questions, or most of them. It should provide systematic answers that address the entire scope of mindreading: the full range of states that get attributed and the full range of contexts or cues on which mental attributions are based. The mental states imputed to others include at least three types: emotions (e.g. fear, anger, disgust), sensations (pain, touch, tickle), and propositional attitudes (belief, desire, intention). Attribution of all such states needs to be covered by an adequate theory. Our own approach will offer plausible answers to all of the foregoing questions. In that sense it constitutes a “full scope” approach. Some of its rivals, by contrast, don’t pass this test of adequacy. The rationality or teleological approach, for example (cf. Dennett, 1987; Gergely, Nadasdy, Csibra, & Biro, 1995; Csibra, Biro, Koos, & Gergely, 2003) seems to lack the resources to explain attributions of sensations or emotions.²

¹ People read their own minds as well as the minds of others. How first-person mindreading is executed is a question of equal importance and difficulty as the third-person mindreading question. For this reason it cannot be addressed within the confines of this chapter, so it is left for another day. (See Goldman, 2006, Chapters 9–10 for an earlier foray into this territory.)

² A standard taxonomy of approaches to mindreading other than simulation theory includes the rationality theory, the child-scientist version of the theory-theory, and the modularity version of the theory-theory.

Levels of mindreading

The general contours of the simulation approach have been laid down by a number of contributors. In the 1980s philosophers advocated simulation (or “replication”) as an alternative to the dominant functionalist, or theory-theory, approach to folk psychology (Gordon, 1986; Heal, 1986; Goldman, 1989). In the 1990s a developmental slant on simulation theory was presented by Harris (1992). Later in the 1990s and 2000s neuroscientific findings steered much of the impetus for simulation theory (Gallese & Goldman, 1998; Currie & Ravenscroft, 2002; Decety & Greze, 2006; Goldman & Sripada, 2005; Gallese, 2007). The present chapter begins by reviewing the original model that focused on the mindreading of propositional attitudes. It then moves to a later variant directed at the attribution of emotions, sensations, and intentional motion. The second half of the chapter examines more recent findings that could play pivotal roles in the ongoing debate.

Many treatments of theory of mind postulate two or more levels, components, or systems of mindreading, and we too offer a bi-level approach (cf. Goldman, 2006). However, not all duplex theories draw the same partitions or have the same rationale. An early two-component architecture similar to one we favor is that of Tager-Flusberg & Sullivan (2000). They distinguish a “social cognitive” component and a “social perceptual” component. Their social-cognitive component features a conceptual understanding of the mind as a representational system, and is highly interactive with other domains such as language. The social-perceptual component is involved in the perception of biological and intentional motion and the recognition of emotion via facial expression. This distinction maps well onto our distinction between “high-level” mindreading of the attitudes vs. “low-level,” or mirror-based, mindreading of non-propositional states. Essentially the same distinction is carved out neurologically by Waytz and Mitchell (2011). Our two methods of mindreading exemplify many of the contrasts that typify the popular dual-systems, or dual-processes, approach in contemporary cognitive science. Low-level mindreading is comparatively fast, stimulus-driven, and automatic; high-level mindreading is comparatively slow, reflective, and controlled.

Apperly advocates a different two-systems approach (Apperly & Butterfill, 2009; Apperly, 2011). Both of his systems concern the propositional attitudes, but two systems are posited by analogy with numerical cognition. One system is characterized as efficient, but inflexible, the other as flexible, but effortful. It is hard to get a firm grip on his systems, however, partly because the account changes significantly between the two publications. The 2009 publication distinguishes between two types of states that are mindread—“registrations” and “beliefs”—but registrations disappear in the 2011 publication.

High-level simulational mindreading

Early formulations of the simulation theory (ST) correspond to what we call high-level mindreading. To pinpoint its most significant features, we contrast it with its perennial foil, theory-theory (TT). We begin with an example:

Shaun just left the house and drove away. I ask you where he is going, and you reply: “He didn’t say. But I know he wants an espresso and thinks that the best espresso is at Sergio’s café. So he probably decided to go to Sergio’s.”

For detailed expositions and critiques of these rivals, see Goldman (2006), Chapters 3–5. Selected problems for some of these rivals are sprinkled throughout this chapter, but length limits preclude detailed treatments.

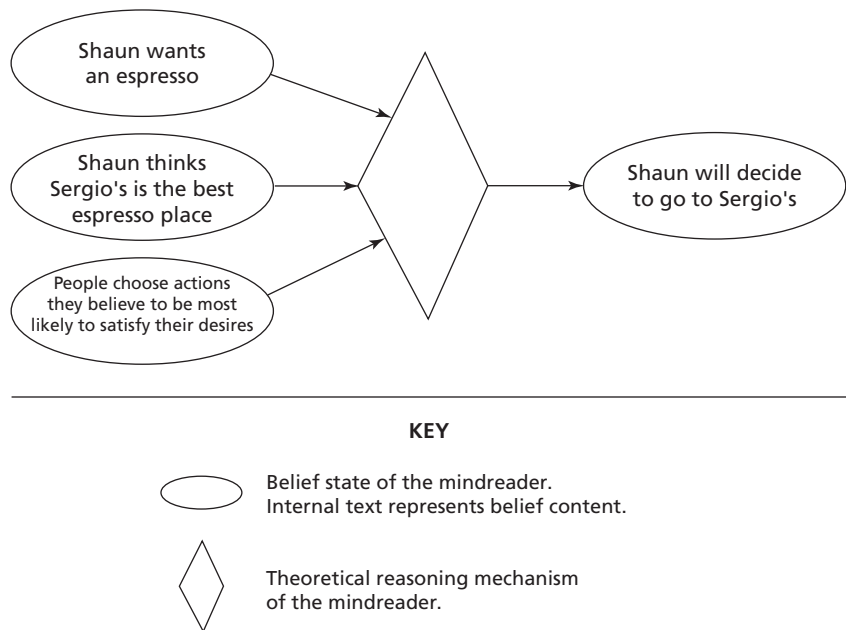


Figure 25.1 TT-type mindreading process. Reproduced from Alvin I. Goldman, *Joint Ventures: Mindreading, Mirroring, and Embodied Cognition*, Figure 5.1 © Oxford University Press, 2013 with permission. For permission to reuse this material, please visit <http://www.oup.co.uk/academic/rights/permissions>.

You have executed a mindreading process, the upshot being the attribution of a certain decision to Shaun. How did you arrive at this? TT would reconstruct your mental process as in Figure 25.1. You start with three beliefs, two specifically about Shaun and one about human psychology in general. All the beliefs are depicted as ovals on the left of Figure 25.1. You believe of Shaun that he wants an espresso and thinks that Sergio's is the best (nearby) espresso place. Your general belief about human psychology is the "theoretical" proposition that people generally choose actions most likely (by their lights) to satisfy their desires. These three "premise" beliefs are fed into your reasoning mechanism, which outputs the conclusion that Shaun decided to go to Sergio's.

Several things about this simple TT story are noteworthy. First, the mindreader's states that do all the "work" in the TT story are belief states, and the only processor used is a theoretical reasoning mechanism. The same would hold of somebody trying to understand and explain the workings of a physical system. Nothing transpires under the theory-oriented account like putting oneself in the target system's shoes. Secondly, the belief states that do the work are structurally rather complex. They are all metarepresentational states, which refer to states of the target that are themselves representational (have content). Shaun is portrayed as having a desire and a belief, each of which is a representational state. Third, under TT, mindreading's aptness for success critically depends on the content of the mindreader's naïve psychological theory. If this theory is ample enough in detail and (approximately) descriptively correct, it may tend to supply fairly accurate mental attributions. But if it is meager or misguided, it will frequently lead the mindreader astray. This will happen especially when the target's mental processes are sophisticated and complex.

One form of TT exploits the adequacy or inadequacy of the mindreader's psychological theory to explain influential patterns of error found in early childhood mindreading, especially errors in (verbal) false-belief tasks. Proponents of this form of TT—so-called "child scientist"

theory-theorists—contend that children gradually refine and improve their theory of mind during their early years, much as adult scientists refine their theories over time. One such refinement is the replacement of a non-representational theory of mind by a representational theory. The later-developing representational theory allows them to conceptualize the possibility of false belief and thereby improve their performance in false-belief tasks between 3 and 4 years of age.

A second form of TT, the modularity theory, denies that children develop a theory of mind by a science-like process. Rather a theory of mind is an innate endowment of one or more dedicated modules (Baron-Cohen, 1995; Leslie, 1994). How, then, might this sort of theory explain comparatively poor performance by 3-year-olds on false-belief tasks? Leslie, German, and Polizzi (2005) introduce an additional, non-modular mechanism, the selection-processor, to account for this phenomenon. The selection processor selects among candidate belief contents in a target agent's mind by inhibiting the default content—namely, the content true of the world—and instead selects an alternative content (which is false of the world). Three-year-olds are weak at this task because their selection processor includes an inhibitory mechanism that hasn't fully matured at three years. Thus, 3-year-olds have a performance problem with false-belief tasks, not a conceptual problem, as child-scientist theory-theorists claim. Despite this difference, both types of TT hold that mind-reading is executed by reliance on a theory of mind, whether an innate theory (embedded in one or more modules) or a gradually developing one.

Could the same tasks be executed in a less informationally demanding manner? Specifically, could they be done with less reliance on refined generalizations about causal connections among mental states? ST takes this tack. It conjectures that mindreaders exploit their own mind as a prototype, or model, of the target's mind. If different minds have the same fundamental processing characteristics, and if the attributor puts her own mind in the same “starting-state” as the target's and lets it be guided by her own cognitive mechanisms, mental mimicry may allow her to determine what the target is going to do. ST embraces this alternative hypothesis, depicted in Figure 25.2.

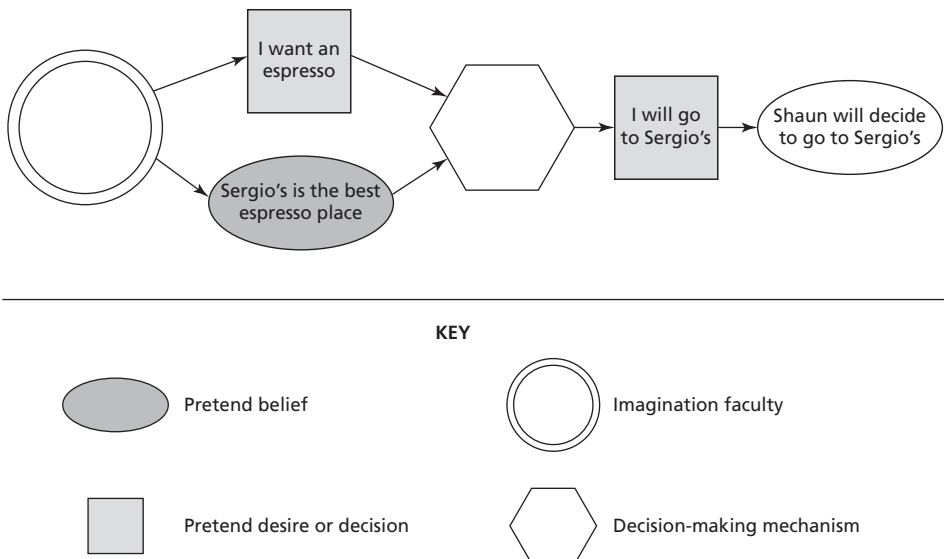


Figure 25.2 High-level ST mindreading process. Reproduced from Alvin I. Goldman, *Joint Ventures: Mindreading, Mirroring, and Embodied Cognition*, Figure 5.1 © Oxford University Press, 2013 with permission. For permission to reuse this material, please visit <http://www.oup.co.uk/academic/rights/permissions>.

A distinctive feature of Figure 25.2 is the presence of shaded shapes, which represent pretend states, i.e. products of pretense or imagination. Imagination is assumed to be a faculty that, when you wish to be in a specified mental state *M*, proceeds to construct an *M*-like state in you. By an “*M*-like state” we mean a state that is (at least functionally) very similar to a genuine state *M*, but would normally be produced by cognitive mechanisms other than imagination (e.g. perception, reasoning, emotion-generation). One crucial similarity between a genuine *M*-state and *M*-like state is that they produce similar output states when fed into a cognitive mechanism, for example, a choice or decision-making mechanism. Figure 25.2 depicts the mindreader as constructing a *pretend* desire (intended to “match” Shaun’s relevant desire or goal) and a pretend belief (intended to match Shaun’s relevant belief). These pretend states are fed into a decision-making mechanism, which operates over these inputs and generates an output state: a decision. This output state is also depicted by shading because it is still under the control of the imagination. Notice that Figure 25.2, unlike Figure 25.1, makes no reference to a factual reasoning mechanism or to a psychological theory of mental processes. The need for theoretical reasoning is replaced in ST by a simulation process, which in this case employs a decision-making mechanism that helps to replicate Shaun’s decision-making process. The simulation routine terminates when the decision-making mechanism outputs a decision. This decision is attributed to Shaun (as shown at the far right of the diagram), the attribution being a genuine (hence unshaded) belief of the mindreader.

Do any interesting predictions flow from the simulationist model? One prediction is that if the mindreader’s imagination performs poorly in constructing the target’s starting state(s), the mindreading routine is not likely to succeed (be accurate). A second prediction requires more ground-laying. Mindreading by simulation runs the risk of letting the mindreader’s own mental states get entangled with the pretend ones. A mindreader, after all, will always have her own “genuine” desires, beliefs, and intentions alongside the pretend ones. These genuine desires and thoughts must be segregated from the pretend ones, an activity that may not be trivial. There is a danger that genuine states will interfere with pretend ones, causing confusion and error. To avoid such entanglement, genuine states must be “quarantined” or “inhibited” to avoid confusion with mimicked states of the target. Thus, intensive use of simulation predicts a high incidence of mindreading error—specifically, egocentric error, reflecting the penetration of the mindreader’s own genuine desires, beliefs, and emotions into the interpersonal tracking process.

Would egocentric errors be equally predicted by TT? Since a theorizing mindreader would also have her own thoughts running on a parallel track with those of the target, doesn’t she face an equal danger of interference? If so, egocentric mindreading errors will not constitute a discriminating test of the rival theories. We argue that the likelihood of interference is higher under ST than TT. Why? Because customary cognitive acts and processes are more similar to—hence easier to confuse with—states of mental mimicry posited by ST than the kinds of cognitive acts and processes posited by TT. Under TT, states deployed during mindreading are exclusively metarepresentational, whereas those deployed during simulation are first-order states. Hence, it should be easier for normal thoughts and plans to mistakenly encroach or insinuate themselves into simulated thoughts and plans than under theory-guided mindreading.

Now, the mindreading literature is replete with reports of egocentric errors, or biases (including, but not restricted to, false-belief attribution errors). Much of it goes under the heading of “curse of knowledge,” a phrase originally introduced in a study of adults who were forewarned that their targets’ knowledge differed from their own, but nonetheless allowed their own knowledge states to seep into attributions to their targets, generating poor task performance (Camerer,

Loewenstein, & Weber, 1989; Nickerson, 1999). The same leakage phenomenon is found in children (Birch & Bloom, 2003, 2004). For the reasons indicated, this lends greater support to ST as compared with TT.³

Although ST easily comports with the observed pattern of egocentric biases, shouldn't it predict much more error than is actually observed? Shouldn't simulation lead to rampant error in virtue of the fact that pretend beliefs, desires, and emotions must surely be different from their genuine counterparts? How can imagination-generated states resemble genuine states so closely that similar decisions or new beliefs get outputted when the pretend vs. genuine states are inputted into similar cognitive mechanisms? Is there really evidence for a tight enough similarity between pretend and genuine states to support high levels of mindreading accuracy? Yes. Cognitive science and neuroscience is replete with evidence that imagination is powerful enough to produce states that closely match their counterparts. This is most thoroughly researched in the domains of visual and motoric imagery. Neuroscientific studies confirm that visual and motor imagery has substantial neurological correspondence with vision and motor execution respectively (Kosslyn, Thompson, & Alpert, 1997; Kosslyn, Pascual-Leone, Felician, & Camposano, 1999; Jeannerod, 2001). Chronometric studies of motor imagery are particularly striking (Decety, Jeannerod, & Preblanc, 1989; Currie & Ravenscroft, 2002, p. 75 ff). Just how powerful and accurate is imagination in non-perceptual and non-motoric cases, however? A recent study described in the section below entitled "The power of imagination" demonstrates the unexpected power of imagination, which should help deflate skepticism about simulation.

A major strand of the ST-TT debate has hinged on the plausibility of the thesis that a complex skill like mindreading is driven by a theory that unfolds during a child's early years. Early defenders of (the child-scientist version of) TT claimed to find evidence that children revise their theory of belief between two and four years of age, yielding mature competence only around four (Wellman, 1990; Perner, 1991; Gopnik & Meltzoff, 1997). The details of this claim, however, were blown out of the water when Onishi & Baillargeon (2005) found false-belief competence (in non-verbal tasks) at 15 months of age. In many parts of cognitive science, however, there is impressive evidence of statistical learning (specifically, Bayesian learning). Does this support a new form of TT as over against the simulation hypothesis?

A study by Baker, Saxe, & Tenenbaum (2009; cf. Baker, Saxe, & Tenenbaum, forthcoming) is a good example of an empirical defense of theory-based mindreading supported by Bayesian inference. They propose "a computational framework based on Bayesian inverse planning for modeling human action understanding ... which represents an intuitive theory of intentional agents' behavior The mental states that caused an agent's behavior are inferred by ... Bayesian inference, integrating the likelihood of the observed actions with the prior over mental states" (2009, p.329). If the cognitive reality of this framework is indeed empirically sustained, as they claim, doesn't it decisively support TT over ST?

Interestingly, Baker and colleagues themselves concede that their findings do not favor TT over ST. The reason is that mindreaders who use Bayesian methods to ascribe mental states to others

³ The reasoning relies on Bayesian conditionalization. When the likelihood of O given H_1 is greater than the likelihood of O given H_2 , observation of O will increase the posterior probability of H_1 more than it increases the posterior probability of H_2 . The present contrast between ST and TT is aimed mainly at the child-scientist version of TT. As reported above, Leslie's form of modularity theory shares with ST a reliance on inhibitory mechanisms.

may simply be running their Bayesian reasoning capacity as a simulation of the target. Thus, as Baker et al. write:

[T]he models we propose here could be sensibly interpreted under either account.... On a simulation account, goal inference is performed by inverting one's own planning process—the planning mechanism used in model-based reinforcement learning—to infer the goals most likely to have generated another agent's observed behavior. (2009, p.347)

Low-level simulational mindreading

The best evidence for low-level simulational mindreading, we submit, is found in research on emotion **mirroring**. First we examine evidence for the existence of mirroring, that is, mental-state contagion. Then we present evidence that mirrored states are used as the causal basis for mindreading.

Both animal and human studies show that the anterior insula is the “gustatory cortex” and primary locus of the primitive distaste response, disgust (Rozin, Haidt, & McCauley, 2000; Phillips, Young, Senior, Brammer, Andrew, Calder, et al., 1997). Against this background, Wicker, Keysers, Plailly, Royet, Gallese, & Rizzolatti (2003) performed a functional imaging study in which participants first viewed movies of other people smelling the contents of a glass (disgusting, pleasant, or neutral) and displaying congruent facial expressions. After first serving in this observer capacity, the same participants then had their own brains scanned while inhaling disgusting or pleasant odorants through a mask on the nose and mouth. The core finding was that the left anterior insula and the right anterior cingulate cortex (ACC) were preferentially activated both during the inhaling of disgusting odorants and during the observation of facial expressions of disgust. Thus, there is indeed mirroring (or contagion) for disgust.

This study presented no evidence concerning the mindreading of disgust (via observed facial expressions). For evidence that mindreading *is* based on mirrored disgust, however, we turn to neuropsychology. Patient NK, studied by Calder et al. (2000), suffered damage to the insula and basal ganglia. In questionnaire responses NK showed himself to be selectively impaired in experiencing disgust. He was also significantly and selectively impaired in recognizing—that is, attributing—disgust. It is hard to explain why NK would have this paired deficit unless the experience of disgust is (normally) causally involved in its attribution. A paired deficit in experience and attribution of disgust seems to be most readily explicable on the assumption that disgust attribution (in observational circumstances) is mediated by its experience. In other words, a normal person uses his intact disgust-experience system to attribute disgust to others. Note that NK was normal with respect to attributing other emotions via observation of facial expressions. Nor was a visual deficit a possible explanation, since NK had the same selective deficit in attributing disgust based on non-verbal *sounds*. Similar findings exist with respect to fear and the amygdala (Goldman & Sripada, 2005; Goldman, 2006: 115 ff).

TT proponents have not offered any systematic account of these findings. One cannot appeal to damage to a hypothesized theorizing system to account for the disgust-attribution impairment, because the relevant patients performed normally when attributing other emotions based on facial or auditory stimuli. Is there a separate theorizing system for each distinct emotion, and was such a theorizing system coincidentally impaired when disgust experience was impaired? Recalling our criteria of adequacy proposed in section 1, the absence of any story of face-based emotion mindreading is a significant count against TT.

Sensations such as pain are another sub-category of low-level mindreading. The most relevant studies here are by Avenanti, Buetti, Galati, & Aglioti (2005), and Avenanti, Paluello, Bufalari, & Aglioti (2006). When a participant experiences pain, motor-evoked potentials (MEPs) elicited by

transcranial magnetic stimulation (TMS) indicate a marked reduction of corticospinal excitability. Avenanti and colleagues found a similar reduction of excitability when participants merely observed someone else receiving a painful stimulus, for example, a sharp needle being pushed into his hand. Thus, there appeared to be mirroring of pain in the observer. Moreover, when Avenanti and colleagues had participants judge the intensity of pain purportedly felt by a model, judgments of sensory pain seemed to be based on the mirroring of pain experienced by the participant.

The conclusion that mirrored pain is the causal basis of pain attribution is clouded a bit, however, by Danziger, Faillernot, & Peyron's (2009) findings from twelve patients with congenital insensitivity to pain. Compared to normal controls in pain recognition tasks, these patients did not differ much in their estimates of the painfulness to other people of various verbally described events. Nor did they differ much from controls in their estimates of degree of pain judged on the basis of facial expression. However, as Carruthers (2011) points out, these individuals with congenital insensitivity to pain may have acquired a different route to pain mindreading than normal people. The findings do not disprove the hypothesis (which Danziger and colleagues embrace) that normal subjects use simulation in reaching their judgments of pain attribution.

Carruthers presses problems for another putative example of low-level simulational mindreading, one concerning face-based mindreading of fear. Adolphs studied a patient, SM, who had a paired-deficit for fear perfectly analogous to the one for disgust displayed by NM. SM, who suffers from bilateral amygdala damage, lacks normal experience of fear and was also selectively impaired in fear attribution. This suggests that the mindreading of fear, like disgust, is ripe for interpretation in simulational terms. However, a later study of SM (Adolphs, Gosselin, Buchanan, Tranel, Schyns, & Damasio, 2005) showed that she was abnormal in scanning her target's eye areas. When she was directed to scan the eyes thoroughly, she improved on fear attribution. Thus, use of fear experience is apparently not strictly necessary for face-based fear attribution, which ostensibly runs counter to a low-level simulational story for fear attribution. However, we can make a similar hypothesis about this case as Carruthers does for those patients congenitally insensitive to pain. Perhaps SM simply developed (under instructions) a skill for face-based mindreading that differs from the simulation heuristic used by normal subjects.

Moreover, other patients with amygdala damage have been studied, with results that support the ST story. Sprengelmeyer, Young, Schroeder, Grossenbacher, Federlein, Buttner, et al. (1999) studied patient NM, who showed selective fear-recognition impairment not only using visual face observation, but also using postural and vocal emotional stimuli. These recognition impairments of NM cannot be explained by appeal to inadequate facial scanning, because the targets' eyes were not visible during the bodily posture task, and played no role in the vocal expression task. So it seems that fear impairment due to amygdala damage does indeed provide a causal explanation of NM's recognition impairment, in conformity with the ST account.

Even if one grants that mental attribution in these cases is caused by the mirroring of others' emotions and sensations, one might balk at the idea that this qualifies as simulation-based attribution. Why does it so qualify? Here is our answer. Consider the diagram in Figure 25.3, where an unshaded shape (on the left side of the diagram, depicting mental states of the target) represents an actual occurrence of disgust and a shaded shape (on the right side of the diagram, depicting states of the observer) represents an observation-induced disgust experience. Just as the Figure 25.2 mindreader imputes a specific decision to her target because she herself "makes" that very decision, so the Figure 25.3 observer undergoes a mirrored experience of disgust, classifies it as an instance of disgust, and projects—i.e. imputes—it to the target. Such a projection of a self-experienced state is a signature of simulational mindreading. Thus, it seems reasonable to regard this as a species of simulation and simulation-based mindreading, even though it is distinguishable from high-level simulational mindreading in some

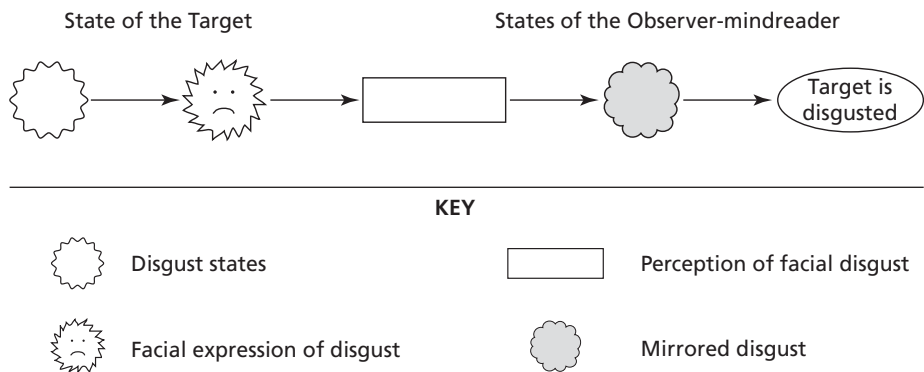


Figure 25.3 Low-level ST mindreading process. Reproduced from Alvin I. Goldman, *Joint Ventures: Mindreading, Mirroring, and Embodied Cognition*, Figure 5.1 © Oxford University Press, 2013 with permission. For permission to reuse this material, please visit <http://www.oup.co.uk/academic/rights/permissions>.

respects (for example, by not being a product of imagination). Both cases involve a process of (genuine or attempted) mental matching between attributor states and target states. The mirror-produced mindreading may be called a single-step simulation process because simulation takes place more or less directly, whereas the decision case is a multi-step simulation process.

Our discussion thus far reviews older evidence pertaining to high- and low-level simulational mindreading. In the remainder of the chapter we adduce more recent lines of evidence and assess their bearing on simulation theory and its competitors.

The power of imagination

As noted under “High-level simulational mindreading,” ST implies that mindreaders’ success in high-level mindreading depends on their ability to enact starting states that sufficiently match those of the target. It follows that if simulational mindreading is to succeed, imagination must be a highly precise mechanism, capable not only of generating suitable pretend states, but of firmly holding their progeny in mind while a multi-step simulational exercise unfolds. What is called for is no minor feat. Is the human imagination powerful enough to meet the challenge?

What exactly does it mean to say that we imagine things from another person’s perspective? In what sense are mindreaders capable of imagining how a target thinks or feels? The sense of imagination we have in mind is a kind of enactment imagination, or E-imagination (Goldman, 2006, Chapter 7). To E-imagine a state is to recreate the feeling of a state, or conjure up what it is like to experience that state—in a sense, to enact that very state. To E-imagine feeling embarrassed involves using one’s imagination to create inside oneself a pretend state that phenomenally feels somewhat like embarrassment. This enactment sense of imagination should be distinguished from another everyday notion of imagination that consists in imagining that such-and-such is the case (as if someone asked you to imagine that you are embarrassed.). This ordinary sense of “imagine” means something like suppose or assume *that* you are embarrassed—which merely requires you to think about or consider a hypothetical situation of embarrassment. It does not require you to conjure up in yourself something resembling the feeling of embarrassment.⁴

⁴ We intend our use of the word ‘imagination’ to be understood in the E-imagination sense.

Inspired by evidence of similarities between perception and mental imagery, researchers recently conducted a study to test the effect of imagined eating on actual subsequent eating (Morewedge, Huh, & Vosgerau, 2010). The results indicate a striking similarity between the states that result from actual eating and those that result from merely imagined eating—namely, both activities result in habituation to the presented stimulus. A series of experiments were conducted to ensure that it was in fact the act of imagining eating that led to a decrease in consumption, and that this decrease in consumption was, indeed, an effect of habituation. The important experiments and results for our purposes are summarized next.

The first experiment was designed to test whether repeatedly imagining consuming a particular food would influence subsequent consumption of that food. Participants were divided into three groups, each of which imagined performing 33 repetitive actions. The control group imagined putting 33 quarters into a laundry machine (an action similar to putting M & M's in one's mouth); the second group imagined putting 30 quarters into a laundry machine and then eating 3 M & M's; and the third group imagined putting 3 quarters into a laundry machine and then eating 30 M & M's. All participants then ate freely from a bowl of M & M's until they indicated that they were finished eating. How much each participant ate from the bowl was measured and compared. The results showed that participants who imagined eating 30 M & M's subsequently ate significantly fewer M & M's than participants in the other groups.

Another experiment tested whether the decrease in consumption was due to habituation or if it was a priming effect resulting from repeated exposure to the stimulus. This time participants either imagined eating 3 or 30 M & M's or they imagined putting 3 or 30 M & M's into a bowl; then, as before, participants ate freely from a bowl of M & M's. Again, results revealed that subjects who imagined eating 30 M & M's ate significantly less than those who only imagined eating 3; but results also showed that subjects who imagined putting 30 M & M's in a bowl ended up eating significantly *more* than subjects who imagined putting only 3 in a bowl. This experiment strongly suggests that not only is priming not the cause of the decrease in consumption, but may even have the opposite effect of increasing subsequent food intake. A further experiment was designed to determine if imaginary eating was habituating people to particular food (causing them to eat less of it), or if it was an overall primed feeling of “fullness” responsible for the decrease in food intake. Here participants imagined eating either 3 or 30 M & M's or cubes of cheese, and then ate freely from a bowl of cheese cubes. The participants who imagined eating 30 cubes ate significantly less than those who imagined eating 3; but participants who imagined eating 30 M & M's did not differ in subsequent cheese consumption from those who imagined eating 3 M & M's. Thus, it seems that the effect of imaginary eating is stimulus specific—providing additional evidence that the reduction in food intake is a result of habituation, and not of priming.

The results of this study are striking. Merely imagining eating can impact how much we actually eat. But how is a study concerning food consumption relevant to mindreading? We argue that this study is easily interpreted as a demonstration of the power of imagination, and to that extent supports our version of ST. In order for the case to be convincing, there are two things that need to be established: (1) the study's use of imaginary eating counts as an instance of imagination in our enactment sense of the word, and (2) the states generated by the imagination really do appropriately resemble their actual counterparts. Our first task is relatively straightforward given the study's experimental design. Participants in the first experiment were asked to repeatedly imagine themselves eating units of food one at a time, not merely to imagine that they had eaten a certain amount of food. Furthermore, results indicated that merely thinking about a particular food repeatedly was not enough. For the habituation effect to occur a person had to actually imagine undergoing the experience of eating a particular food. However, this is just how we have

characterized an act of enactment imagination: as an attempt to re-enact or re-experience a particular feeling or state.

What about our second task? Have Morewedge and colleagues shown that imagining eating is capable of producing accurate pretend states, similar to those that result from actually eating? When a person actually eats a particular food, they gradually habituate. Their desire to eat the food, along with the motivation to obtain it, gradually decreases. If presented with a different food, however, the person's desire and motivation recover. This implies that habituation effects are stimulus specific (Epstein, Saad, Handley, Roemmich, Hawk, & McSweeney, 2003). If this is what happens when we actually eat, then something sufficiently similar to habituation should result when we repeatedly imagine ourselves eating: additionally, it should be the case that if we imagine eating something only a few times we do not habituate. The results of the Morewedge et al. study clearly demonstrate that imagined eating results in habituation, similar to when a person has actually eaten. Furthermore, given that habituation effects are stimulus specific the imagined consumption of a particular food should cause a person to habituate to that food only. As the study demonstrates, this is exactly what happens.

Imagined consumption is a clear instance of enactment imagination as well as of the resemblance that can obtain between imagination-induced states and their genuine counterparts. But imaginary eating is not a case of mindreading. Can more be done to make the connection between this research and simulational mindreading clear? After all, if ST is right, mindreaders use their imagination in tasks involving a variety of mental states. So we need to establish that imagination can produce pretend states that closely resemble actual states across a respectable spectrum of cases. We maintain that research on imagined consumption gives us reason to think the imagination has this capacity.

According to Morewedge et al., "Habituation to food occurs too quickly for it to result from digestive feedback, so it is commonly thought to occur as a result of top-down cognitive processes (such as beliefs, memories, or expectations) or pre-ingestive sensory factors (such as texture or smell)" (2010, p.1531). This study demonstrates that habituation can occur as a result of imagination alone, without any influence from sensory information. This is significant because habituation is a very general phenomenon, specific neither to food nor to eating. Research indicates that we habituate to a wide range of complex emotions, attitudes, feelings, and moods: from states of happiness and love to states of fear and anxiety (Solomon, 1980). This study confirms that the power of imagination could be very general indeed.⁵

Mindreading acquisition

Recall that our second criterion of adequacy requires a comprehensive theory of mindreading to give a plausible, empirically sustainable story about how the mindreading capacity is acquired. Past research on the time-course of childhood mentalizing had important implications for theories of how mindreading is acquired, such as the child-scientist approach to theory-theory. This

⁵ Additional confirmatory evidence comes from research on memory distortion involving imagination (Schacter, Guerin, & St Jacques, 2011). A study by Mazzoni & Memon (2003) indicated that the strength of subjects' beliefs that events occurred increased more when they imagined events than when they simply read about them. Nash et al. (2009) showed that imagining that one has performed an act produces about as many false memories of actually having done it as viewing a doctored video that suggests that one did perform the act.

approach claims that mental-state attributions are driven by naive psychological theories that are initiated and gradually revised in the early years. This claim has been increasingly undercut, however, by recent research revealing sensitivity to false beliefs even in pre-verbal infants (e.g. Onishi & Baillargeon, 2005).

How does the simulation theory fit with such evidence? What is ST's position on the acquisition of the mentalizing capacity? ST does not take a firm stance vis-à-vis nativism. It is prepared to "go with the flow" of evidence. For example, it is prepared to say that the processes or methods of mindreading (or dispositions to use such processes) are part of our native endowment. It might be more skeptical about claims that particular mental-state concepts (belief, desire, pleasure, etc.) are all innate. It is prepared, however, to accommodate the former type of nativism if empirical studies provide warrant for this position. ST's theoretical apparatus does not preclude such strands of nativism. Indeed, one might say something stronger from the vantage-point of ST. If imagination is an innate capacity, perhaps young infants automatically compute imaginary states for people around them. Then we would expect the practice of generating imaginary states to be no more cognitively demanding than one's own largely automatic production of mental states. This section will discuss how well such expectations comport with recent evidence in developmental psychology. Our primary focus is a compelling new study conducted by Kovacs, Teglas, & Endress, (2010) plus the simulational hypothesis we claim to be consistent with this study.⁶

Unlike standard false-belief tasks, this study was designed to investigate mindreading mechanisms *implicitly*—making no reference to others' beliefs, and not requiring any behavioral predictions based on others' beliefs. The study had two components: one testing the reaction time (RT) of adult participants and the other measuring the looking time (LT) of 7-month-old infants. Participants watched a series of short movies involving an animated agent, a ball, and a table with an occluder. At the beginning of each movie the animated agent entered the scene and placed the ball on the table in front of the occluder.⁷ The ball then rolled behind the occluder. At this point, depending on the experiment, the ball either stayed in place or rolled off the screen. Then the agent left the scene. The ball's final location and the time the agent left the scene were varied, such that the agent would have a true belief about the ball's location if he left after the ball reached its final location and a false belief if he left before. The critical variables involved the participant's beliefs about the ball's presence or absence and the agent's "beliefs," such that the participant, the agent, both, or neither could believe the ball was behind the occluder (Kovacs, Teglas, & Endress 2010, p. 1831). At the end of the movie, the agent re-entered the scene and the occluder was lowered, revealing the ball to be either present or absent. Adult participants were instructed to press a button as soon as they detected the ball. Their RTs and the infants' LTs were measured in each of the four conditions.

The experimental conditions, for both adults and infants, were compared to a baseline condition where neither the participants nor the agent believed the ball to be behind the occluder.⁸ The most important experiment with adult participants was one in which only the agent believed the ball

⁶ Although we focus on the results of the Kovacs et al. study, we find the wording of their study potentially misleading and worry that it may not convey the researchers' true intentions. Because of this, we will outline the results as reported in the study, but the conclusions drawn will be our own and are not intended to match those of the authors.

⁷ 'Agent' always refers to the animated character in the film and 'participant' refers to the adult or infant participant in the study.

⁸ However, recall that in none of the conditions were the agent's beliefs relevant to the task.

to be behind the occluder. Results indicated that participants' RTs were faster in this case than in the baseline condition, despite no difference in the participants' beliefs in either condition. Kovacs et al. take this result to demonstrate that the participants not only automatically computed the agents' beliefs, but that these beliefs influenced the participants' behavior, despite the agents' beliefs being inconsistent with their own (1832). Additionally, the participants' RTs did not significantly differ when only the agent believed the ball to be behind the occluder and when they themselves believed it to be. Kovacs et al. further conclude: "Thus both types of belief representations speeded up the participants' RTs to similar extents, a result consistent with the view that the agent's beliefs are stored similarly to participants' own representations about the environment" (2010, p.1832).

The crucial results with the infant participants similarly involved a comparison between the infants' LTs in two conditions: one in which only the agent believed the ball to be behind the occluder and the other in which neither the infant nor the agent believed the ball to be there. When no ball appeared behind the occluder, the infants looked longer (indicating their "surprise" by the outcome) in the condition where only the agent believed, or expected, the ball to be there. Again, this suggests not only that the infants computed the agent's belief, but also that this belief influenced the infants' behavior despite conflicting with their own (genuine) beliefs. It is similarly interesting that with both adults and infants, very similar results obtained even when the agent did not return to the scene and thus was not present when the occlusion was lowered. Infants and adults seemed to compute and maintain the agent's beliefs even when the agent was no longer present.

What do these results mean for the study of mindreading? More specifically, how do they fit with what ST says about mindreading? Concerning the question of acquisition, the infant results are of primary interest. How do they fit with theory-driven vs. simulation-driven mindreading processes depicted in Figures 25.1 and 25.2 respectively? A Figure 25.1-type story would say that at 7 months of age infants not only compute the beliefs of other agents, but that these computations are based on the infants' beliefs about the beliefs of the agent. In other words, TT-type explanations rely on the infants' possession of relatively complex metarepresentational states, plus their possession of some body of psychological laws or generalizations.⁹ Thus, TT's approach to mindreading is information rich, and requires a degree of cognitive or informational sophistication that one may be hesitant to attribute to 7-month-old infants.

By contrast, the Figure 25.2-type story suggests that the same sort of tracking of the agent's thoughts has another, simpler interpretation. ST implies that infants track an agent's perspective in the same way they maintain their own perspective. Just as infants have their own current representations of the environment, they also track other possible states of the environment. This sort of explanation, in contrast to Figure 25.1-type theories, is an information-poor approach, because it does not attribute to the infants any additional theoretical knowledge or metarepresentational states. To perform perspective computations, ST only requires that infants possess states with object-level representational content—information about the way the world seems from the shoes of the agent. This means that infants may track the content of an agent's belief (possible states of the environment) without encoding anything concerning his beliefs or other mental states.

Given what we have said so far, ST is in as good a position as TT to account for the Kovacs study. Might there be reasons to think it may be in a better position to explain its findings? We argue that there are. First of all, TT has to say that pre-verbal infants compute metarepresentations. Is it psychologically plausible to impute such cognitively complex mental states to infants? Would it not be

⁹ Depending on the particular TT-type approach we are discussing, such theories may also require that infants possess other complex theoretical beliefs about human psychology.

preferable, if possible, to account for the infants' behavior without attributing to them such extra computational work or informational baggage? If so, then the ST explanation is clearly preferable, because it accounts for the evidence without positing the extra complexity of metarepresentational states or a body of psychological generalizations.

Concerning the question of acquisition, there are other reasons to think the Kovacs study supports a simulation story about mindreading. What this study shows, we have claimed, is that 7-month-old infants generate representations of the world that reflect another person's perspective—but to represent the world as it seems to another person just *is* to use one's imagination.¹⁰ Although it seems unlikely that at 7 months infants engage in explicit acts of mindreading (i.e. attribution of mental states to others), they certainly appear to engage in mindreading-like activity; furthermore, this mindreading-like activity involves use of their imagination. This means that before they ever engage in a single act of mindreading, infants are already experienced imaginers. By the time they get to the point of attributing mental states to other people, they have spent years spontaneously and automatically imagining the world from other people's perspectives.

The neural basis of mindreading

Now we apply the third question of adequacy to our version of simulation theory: Is this theory neurally plausible, given available empirical evidence? One issue is whether recent evidence from cognitive neuroscience supports (or is consistent with) the claim that simulation is a common method, if not the predominant method, of mindreading. A second issue is whether neuroscience supports our specific version of ST, i.e. a bi-level or duplex version of ST. Because neuroscientific evidence was already adduced in support of the existence of mirroring and the grounds for linking it to ST, we won't say more about the first issue. We shall concentrate on the second.

Waytz & Mitchell (2011) present the neuroscientific case for a duplex model of simulational mindreading as follows. First, they review the extensive evidence of multiple mirroring phenomena, sometimes referred to as "shared neural representations." These include regions in the inferior frontal cortex and superior parietal lobe (i.e. the parieto-frontal circuit) which are involved in the production and observation of goal-directed motor action.¹¹ They also include a wide range of regions for the mirroring of pain, touch, disgust, and fear (cf. Rizzolatti & Sinigaglia, 2008). Networks in these areas are what we treated under the heading of low-level simulation.

Another set of brain regions has been identified, however, that serves as a substrate for what Buckner & Carroll (2007) call self-projection. These regions, known collectively as the "default network" (Raichle, MacLeod, Snyder, Powers, Gusnard, & Shulman, 2001), consist of the medial prefrontal cortex, precuneus and posterior cingulate, and lateral parietal cortex. The default network

¹⁰ This is where our conclusions may come apart from those drawn by Kovacs et al. While the conclusions drawn in the study seem to claim that 7-month-old infants are representing the beliefs of the agent alongside their own beliefs; we claim that the results of this study only demonstrate that infants are generating representations of the world that reflect what the target's or agent's beliefs would be (not that they represent them *as* beliefs). Furthermore, the study itself makes no mention of the imagination—rather, the results demonstrate that infants engage in an activity that epitomizes our conception of the use of imagination in simulation.

¹¹ The mirroring theory has characteristically claimed that mirroring is used to understand the actions of others. There is continuing debate, however, over which specific brain networks comprise the action-observation system, and how exactly they function. For example, Kilmer (2011) defends a two-pathway model of action understanding, featuring a dorsal pathway in addition to the initially discovered ventral pathway.

has repeatedly been linked to tasks in which people imagine experiencing fictitious events, consider the possibility of experiencing specified events in the future, or recall their experiences from the past. It has also been reported by studies in which participants contemplated other people's mental states (Frith & Frith, 2003). Thus, the default network is an excellent candidate for the neural substrate of high-level simulational mindreading. Moreover, this network seems to be quite different (non-identical) from any of the circuits or processes involved in mirroring.

Finally, Waytz & Mitchell point out that there are dissociable functions of mirroring and self-projection. Perceivers mirror only when they see or hear another person's physical actions, observe an emotional expression, or witness a painful situation such as a needle penetrating a hand. But mindreading also occurs when subjects represent targets who are not immediately present, and hence are not observable. Such targets include fictitious individuals or individuals known only by description, where no observable cues are available. Waytz & Mitchell consider this a demonstration of a dissociation between mirroring and self-projection. They cite a mentalizing study by Zaki, Hennigan, Weber, & Ochsner (2010) in which participants inferred a target's emotional state under three conditions: during perceptually cued trials, during context-only trials, and when participants had both perceptual and contextual information. Consistent with the proposed division of labor between two systems of mentalizing, they found that perceptual cues tended to elicit stronger activation in mirror-related brain regions (the fronto-parietal circuit) whereas contextual cues engaged the default network.

Lombardo, Chakrabarti, Bullmore, Wheelwright, Sadek, Suckling, et al. (2010) take a somewhat different perspective on the two-systems approach based on their finding of functional connectivity patterns during mentalizing of both self and other. They don't deny the dissociability claim of Waytz & Mitchell, but they argue that the functional connectivity patterns revealed in their studies support a slightly different picture than the one offered here. Indeed, they advance the thesis that some aspects of both high-level and low-level social cognitive processes are "grounded" within a framework of embodied cognition. We don't believe that there are fundamental differences between their view and ours. At any rate, we find no reason to disagree with a very similar picture presented by Zaki & Ochsner (2012), who also stress functional connectivity between the two systems during experiences of empathy. As Zaki & Ochsner express it, "naturalistic" (i.e. ecologically valid) situations involve many dynamic social cues (featuring both sensorimotor and contextual information), and such cues unsurprisingly generate dynamical neural interactions among simpler processes (low-level and high-level processes). These more complex processes could not be understood, Zaki & Ochsner acknowledge, without a prior understanding of the simpler processes in isolation, which are coupled during complex social tasks (2012: 678). By our lights, this is a reasonably clear recognition that there *are* simpler processes, which we take to be the low-level and high-level families of processes of our model.¹² It is the existence and distinctness of these "simpler" processes that comprise the core thesis defended in this section.

ST's "mesh" with evolutionary theory

We turn finally (and briefly) to the fourth question of section 1—"Does our theory mesh with successful theories in other cognitive domains and with plausible accounts of the architecture and

¹² One non-trivial point of difference, however, is that Zaki and Ochsner identify the higher-level processes as "mentalizing" processes, implying that lower-level ("shared representation") processes are not involved in mentalizing. By contrast, we claim that the latter also serve as a causal basis of mentalizing.

evolution of cognition?” Do successful theories of other parts of cognition invoke similar explanatory faculties or processes, and does a reasonable account of brain evolution find a natural home for simulationist stories of mindreading?

Begin with ST’s account of high-level mindreading, in which imagination occupies a central role. It makes good sense, we submit, to assign a pivotal role to imagination because this faculty has demonstrated its power and versatility in many other domains of cognition. Its robust power and versatility are amply exhibited in such diverse phenomena as visual and motor imagery, the planning of action sequences, and the reduction of food consumption. With respect to low-level mindreading, the discovery of mirror neurons and mirror systems has revolutionized research and thinking about many aspects of low-level cognition (Rizzolatti & Craighero, 2004; Gallese, Keysers, & Rizzolatti, 2004). Contemporary social neuroscience is replete with new insights related to mirroring. A primitive kind of mindreading based on mirroring is a good fit with much of this literature.

ST also comports well with current understandings of brain evolution. As Anderson (2010) tells the story, it is very common for neural circuits originally established for one purpose to be exapted—that is, exploited, recycled, redeployed—during evolution and put to different uses, without necessarily losing their original functions. Nature has a pervasive strategy of opportunistically exploiting existing neural hardware to solve new problems—or to create new solutions to old problems. Creating whole brain structures *de novo* in order to tackle problems would be expensive. Instead, nature prefers a redeployment strategy. This idea meshes well with ST’s story of low-level mindreading. So, for example, suppose that nature had earlier hit upon the strategy of devising mechanisms by which shared representations are generated in the heads of two interacting individuals. A mental representation (or event) in one individual’s brain leads to the generation of a matching representation (or event) in an observer. Once this kind of interpersonal transmission mechanism has evolved, members of the species can secure valuable information by piggy-backing a mental attribution mechanism on top of the shared-representation, or mirroring, mechanism. This is a cheap way to create a reliable mindreading device. It would be unsurprising if something like this evolutionary story were true.

This is what we mean in saying that ST “meshes” well with what is known, or reasonably believed, about brain evolution. According to many philosophers of science, consilience with existing theory is one form of evidence for a new theory. Thus, another chunk of evidential support is added in favor of ST, on top of the more direct kinds of evidence presented in preceding sections.

References

- Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature* 433: 68–72.
- Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences* 33: 245–66.
- Apperly, I. A. (2011). *Mindreaders: The Cognitive Basis of Theory of Mind*. Hove: Psychology Press.
- Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review* 116(4): 953–70.
- Avenanti, A., Buetti, D., Galati, G., & Aglioti, S. M. (2005). Transcranial magnetic stimulation highlights the sensorimotor side of empathy for pain. *Nature Neuroscience* 8: 955–60.
- Avenanti, A., Paluello, I. M., Bufalari, J., & Aglioti, S. M. (2006). Stimulus-driven modulation of motor-evoked potentials during observation of others’ pain. *NeuroImage* 32: 316–24.
- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition* 113: 329–49.

- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2012). Bayesian theory of mind: Modeling joint belief-desire attribution. In: *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 2469–74).
- Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge: MIT Press.
- Birch, S. A. J., & Bloom, P. (2003). Children are cursed: an asymmetric bias in mental-state attribution. *Psychological Science* 14: 283–6.
- Birch, S. A. J., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences* 8: 255–60.
- Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences* 11: 49–57.
- Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience* 3: 1077–1078.
- Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: an experimental analysis. *Journal of Political Economy* 97: 1232–54.
- Carruthers, P. (2011). *Opacity of Mind*. Oxford: Oxford University Press.
- Csibra, G., Biro, S., Koos, O., & Gergely, G. (2003). One-year old infants use teleological representations of actions productively. *Cognitive Science* 27: 111–33.
- Currie, G., & Ravenscroft, I. (2002). *Recreative Minds*. Oxford: Oxford University Press.
- Danziger, N., Faillernot, I., & Peyron, R. (2009). Can we share a pain we never felt? Neural correlates of empathy in patients with congenital insensitivity to pain. *Neuron* 61: 203–12.
- Decety, J., Jeannerod, M., & Preblanc, C. (1989). The timing of mentally represented actions. *Behavioral and Brain Research* 34: 35–42.
- Decety, J., & Greze, J. (2006). The power of simulation: Imagining one's own and other's behavior. *Brain Research* 1079: 4–14.
- Dennett, D. (1987). *The Intentional Stance*. Cambridge: MIT Press.
- Epstein, L. H., Saad, F. G., Handley, E. A., Roemmich, J. N., Hawk, L. W., & McSweeney, F. K. (2003). Habituation of salivation and motivated responding for food in children. *Appetite* 41(3): 283–9.
- Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transaction of the Royal Society of London, Series B: Biological Sciences* 459: 358.
- Gallese, V. (2007). Before and below “theory of mind”: embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of Royal Society B: Biology* 362(1480): 659–69.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences* 2: 493–501.
- Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences* 8: 396–403.
- Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition* 56: 165–93.
- Goldman, A. I. (1989). Interpretation psychologized. *Mind and Language* 4: 161–85.
- Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, & Neuroscience of Mindreading*. New York: Oxford University Press.
- Goldman, A. I., & Sripada, C. (2005). Simulationist models of face-based emotion recognition. *Cognition* 94: 193–213.
- Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge: MIT Press.
- Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language* 1: 158–71.
- Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind and Language* 7: 120–44.
- Heal, J. (1986). Replication and functionalism. In: J. Butterfield (Ed.), *Language, Mind, & Logic* (pp. 135–50). Cambridge: Cambridge University Press.

- Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage* 14: S103–9.
- Kilmer, J. M. (2011). More than one pathway to action understanding. *Trends in Cognitive Sciences* 15(8): 352–7.
- Kosslyn, S. M., Pascual-Leone, A., Felician, O., & Camposano, S. (1999). The role of area 17 in visual imagery: Convergent evidence from PET and from rTMS. *Science* 284: 167–70.
- Kosslyn, S. M., Thompson, W. L., & Alpert, N. M. (1997). Neural systems shared by visual imagery and visual perception: a positron emission tomography study. *NeuroImage* 6: 320–34.
- Kovacs, A. M., Teglas, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science* 330: 1830–4.
- Leslie, A. (1994). Pretending and believing: Issues in the theory of ToMM. *Cognition* 50: 211–38.
- Leslie, A., German, T., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology* 50: 45–85.
- Lombardo, M. V., Chakraborti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., & Baron-Cohen, S. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience* 22(7): 1623–33.
- Mazzoni, G., & Memon, A. (2003). Imagination can create false autobiographical memories. *Psychological Science* 14: 186–8.
- Morewedge, C. K., Huh, Y. E., & Vosgerau, J. (2010). Thought for food: Imagined consumption reduces actual consumption. *Science* 330: 1530–3.
- Nash, R. A., Kimberly, A. W., & Lindsay, D. S. (2009). Digitally manipulating memory: Effects of doctored videos and imagination in distorting beliefs and memories. *Memory and Cognition* 37(4): 414–24.
- Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: imputing one's own knowledge to others. *Psychological Bulletin* 125: 737–59.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science* 308: 255–8.
- Perner, J. (1991). *Understanding the Representational Mind*. Cambridge: MIT Press.
- Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., Bullmore, E. T., Perrett, D. I., Rowland, D., Williams, S. C. R., Gray, J. A., & David, S. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature* 389: 495–8.
- Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences USA* 98: 676–82.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience* 27: 169–92.
- Rizzolatti, G., & Sinigaglia (2008). *Mirrors in the Brain: How Our Minds Share Actions and Emotions*. Oxford: Oxford University Press.
- Rozin, P., Haidt, J., & McCauley, C. (2000). Disgust. In: M. Lewis and J. Haviland (Eds), *Handbook of Emotions* (pp. 575–94). New York: Guilford Press.
- Schacter, D. L., Guerin, S. A., & St Jacques, P. L. (2011). Memory distortion: an adaptive perspective. *Trends in Cognitive Sciences* 15(10): 467–474.
- Solomon, R. L. (1980). The opponent-process theory of acquired motivation: The costs of pleasure and the benefits of pain. *American Psychologist* 35(8): 691–712.
- Sprengelmeyer, R., Young, A. W., Schroeder, U., Grossenbacher, P. G., Federlein, J., Buttner, T., & Przuntek, H. (1999). Knowing no fear. *Proceedings of the Royal Society, B: Biology* 266: 2451–6.
- Tager-Flusberg, H., & Sullivan, K. (2000). A componential view of theory of mind: evidence from Williams Syndrome. *Cognition* 76: 59–89.
- Watz, A., & Mitchell, J. P. (2011). Two mechanisms for simulating other minds: Dissociations between mirroring and self-projection. *Current Directions in Psychological Science* 20(3): 197–200.

Wellman, H. (1990). *The Child's Theory of Mind*. Cambridge: MIT Press.

Wicker, B., Keysers, C., Plailly, J., Royet, J-P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust . *Neuron* 40: 655–64.

Zaki, J., Hennigan, K., Weber, J., & Ochsner, K. N. (2010). Social cognitive conflict resolution: Contributions of domain-general and domain-specific neural systems. *Journal of Neuroscience* 30: 8481–8.

Zaki, J., & Ochsner, K. N. (2012). The neuroscience of empathy: progress, pitfalls and promise. *Nature Neuroscience* 15(5): 675–80.

Mindreading the self

Peter Carruthers

This chapter contrasts two different kinds of account of our knowledge of our own thoughts. According to standard theories, self-knowledge of at least a subset of thoughts is direct and non-interpretive. According to the alternative, which will be elaborated and defended here, self-knowledge results from turning our mindreading capacities on ourselves, relying on the same sensory channels that we employ for other-knowledge and utilizing many of the same sensory cues.

Introduction

Philosophers have traditionally assumed that self-knowledge is special. Knowledge of one's own thoughts, in particular (one's beliefs, judgments, desires, hopes, fears, decisions, and intentions) is supposed to be especially intimate, direct, and reliable. Indeed, Descartes (1641) famously believed that one's knowledge of one's own current thoughts is infallible (one cannot be mistaken about them), and that those thoughts themselves are self-presenting (to have them is to have infallible knowledge of them). Nor was Descartes by any means alone in holding such beliefs. Similar views were endorsed by Aristotle (see Caston, 2002), Augustine (see Bolyard, 2009; Mendelson, 2009), Locke (1690), and many others. While philosophers today don't endorse anything so extreme, almost all hold that knowledge of at least a subset of one's own thoughts is authoritative (incapable of being challenged by others) and privileged (arrived at in a special way that isn't available to others). Indeed, similar views are common even among cognitive scientists, especially those who believe that third-person mindreading capacities are grounded in first-person awareness (Gallese & Goldman, 1998; Rizzolatti, Fogassi, L., & Gallese, 2001; Goldman, 2006; Meltzoff & Brooks, 2008). In fact, some sort of tacit commitment to the special nature of self-knowledge has a strong claim to be a human universal. For although no work has been done on the subject by anthropologists in small-scale societies, it seems that such views have been endorsed across time and place (whether tacitly or explicitly) whenever people have reflected and written on the question (Carruthers, 2011).

The present chapter will suggest that this widespread view is radically mistaken. Far from being special, self-knowledge results from turning our mindreading abilities on ourselves. The same mental faculty that evolved for reading the minds of others and negotiating the social world gets turned toward the self, issuing in knowledge of our own thoughts (although often, also, in false beliefs about them). On this view, the mindreading faculty is arranged as one of the consumer systems for "globally broadcast" attended perceptual information (in the sense of Baars, 1988, 1997), for of course mindreading would need to have access to such information in order to perform its primary function. Plainly, attributing thoughts to other people requires observations of their behavior and physical circumstances. Self-knowledge can then rely on anything that is accessible through these same sensory channels, including one's own behavior and context, but also one's own visual imagery, inner speech, felt affect, and other forms of sensory experience. For imagery utilizes

the same mechanisms as does perception, and is globally broadcast in the same manner (Kosslyn, 1994; Kosslyn & Thompson, 2003). So while knowledge of our own sensory states is direct, knowledge of our own thoughts is just as interpretive in nature as knowledge of the thoughts of others, and relies on many of the same kinds of sensory cue.

Our discussion of these contrasting accounts will proceed as follows. In “Confabulation and dual-method theories,” evidence of confabulation in attributing thoughts to ourselves will be discussed, providing one major strand of support for the views just outlined. This section will also consider the defensive moves that are available to defenders of the special character of self-knowledge. Then in “The interpretive sensory-access account” the interpretive sensory-access account of self-knowledge will be elaborated in somewhat more detail. “Dissociation data” considers potential dissociation evidence from schizophrenia and autism. “Brain imaging evidence” discusses some of the brain-imaging evidence that bears on the issue.

Confabulation and dual-method theories

More than half a century of careful research in social psychology has produced voluminous evidence that people will often confabulate about their own current or very recently past thoughts. That is, they issue reports of their current or recent thoughts that are manifestly false, but seemingly without any awareness of the falsity of their claims. Moreover, in many cases these reports are just the ones that third-party observers with access to the same information would attribute to the subjects, suggesting that in these instances, at least, self-attributions of thoughts result from turning one’s mindreading abilities on oneself. Philosophers wishing to defend anything resembling the traditional view of self-knowledge have been forced to embrace dual-method accounts as a result of this data (Nichols & Stich, 2003; Goldman, 2006). They claim that sometimes we turn our mindreading abilities on ourselves (often resulting in confabulation), but on other occasions we have access to our own thoughts that is direct and non-interpretive. The present section will sketch some examples of confabulation, before evaluating the dual-method response.

Confabulation data

A significant portion of the evidence has been collected by those working within the “self-perception” framework initiated by Bem (1967). For example, Wells & Petty (1980) found that nodding one’s head while listening to a message on a tape significantly increases people’s expressed agreement with the message thereafter, while shaking one’s head while listening significantly decreases agreement. (Subjects were told that they were testing how well the headphones stay on people’s heads, and that the message was incidental to the purpose of the experiment.) It seems that subjects interpret their own behavior as agreement or disagreement with the message, and adjust their reports of their own degree of belief in the subject of the message accordingly.

Briñol & Petty (2003) replicated this result, and were able to demonstrate that it is not a consequence of priming or positive mood caused by the head movements. They varied the persuasiveness of the message, finding that when the message is persuasive the original result replicates, whereas when the message is *unpersuasive* the opposite occurs: those nodding their heads agree with the message even less, while those shaking their heads agree with it more. The experimenters were able to show that subjects interpret their head movements as agreeing or disagreeing with their own internal reactions while listening to the message (such as saying to oneself, “What an idiot!”), and they were able to find no evidence of mood changes.

Brinöl and Petty (2003) also conducted a separate experiment in which subjects had to write three statements about themselves that might impact their careers, writing either with their right or their left hands. They were then asked for their degree of confidence in the statements that they had written. Right-handers who wrote with their left hands expressed significantly lower confidence, presumably because the shaky writing is interpreted by the mindreading system as a sign of hesitancy and uncertainty. Indeed, third parties who looked only at the written statements and were asked about the degree of confidence of the writer showed exactly the same effect.

These data are consistent with a mixed account, according to which reports of one's thoughts can be influenced by mindreading while *also* depending on some privileged channel of information. But many other results in the literature cannot easily be interpreted in this way. For example, Wegner & Wheatley (1999) asked subjects to report on their intentions in experiments in which (they believed) they were jointly controlling the cursor on a computer screen with another subject (who was in fact a confederate of the experimenters). Immediately after each trial they were asked to record the extent to which they had intended the final position of the cursor on a 100-point scale ranging from 1 ("I allowed the stop to happen") to 100 ("I intended to make the stop"). In one condition the confederate was told to play no part in the movement of the cursor, in fact giving subjects complete control. On average people still rated their degree of intent at only 56, just above the mid-point. Presumably, they made the reasonable assumption that control would be shared and therefore anchored on the mid-point of the scale, only adjusting upwards slightly in conditions in which they in fact had complete control (perhaps being sensitive to the presence of less resistance on the computer-mouse than they had expected).

This is already a remarkable result. For we can assume that the subjects made a decision to stop just prior to the time when they did (since the confederate played no role). We can also assume that they would have been paying close attention to their states of intending, since they knew that they would need to report on them immediately thereafter. But if their own decisions were directly available to them, then one would predict that they should have had a powerful sense of causality in these circumstances. For we know that in general temporally-contiguous events give people a strong sense of causation (McCloskey, Colebatch, Potter, & Burke, 1983; Young, 1995). The absence of any such effect speaks powerfully against the idea of direct introspective access to intentions.

In other conditions the confederate was instructed via headphones in such a way as to bring the cursor to a halt next to a particular type of object depicted on the screen (such as a beach ball), in circumstances in which the subject would hear the name of that type of object (ostensibly as a distracter). When the word was heard many seconds in advance of the stop, subjects on average scored the degree to which they had intended the stop at 45, presumably because they were sensitive to some resistance in the movements of the mouse in conditions in which the confederate in fact had ultimate control. But when the word was heard just before the stop, they scored their degree of intent at over 60. Presumably, their mindreading systems interpreted the coincidence of stopping near the object that had just been named as evidence of an intention to stop at that point.

Let me finish this brief sampling of confabulation data with some discussion of the "dissonance" tradition in social psychology, from which hundreds of supporting references could be provided. In a typical type of experiment, subjects will be induced to write an essay arguing for a conclusion that is the contrary of what they believe. In one condition, subjects may be led to think that they have little choice about doing so (for example, the experimenter might emphasize that they have previously agreed to participate in the experiment). In the other condition, subjects are led to think that they have freely chosen to write the essay (perhaps by signing a consent form on top of the essay-sheet that reads, "I freely agree to participate in this experiment.")

The normal finding in such experiments is that subjects in the free-choice condition (and only in the free-choice condition) change their reported attitudes on the subject-matter of the essay. This happens, although there are typically no differences in the quality of the arguments produced in the two conditions. If subjects in the free-choice condition have previously been strongly opposed to a rise in university tuition costs, for example (either measured in an unrelated survey some weeks before the experiment, or by assumption, since almost all people in the subject pool have similar attitudes), then following the experiment they might express only weak opposition or perhaps even positive support for the proposed increase. Such effects are generally robust and highly significant, even on matters that the subjects rate as important to them, and the changes in reported attitude are often quite large.

We know that freely undertaken counter-attitudinal advocacy gives rise to negatively valenced states of arousal, which dissipate as soon as subjects express an attitude that is more consistent with their advocacy (Elliot & Devine, 1994). Indeed, even *pro*-attitudinal advocacy will give rise to changes in expressed attitude in circumstances where subjects are induced to believe that their honest advocacy will turn out to have bad consequences (Scher & Cooper, 1989). In circumstances where subjects are offered a variety of methods for making themselves feel better about what they have done (an attitude questionnaire, a question about their degree of responsibility, and a question about the importance of the topic), they will use whatever method is offered to them first (Simon, Greenberg, & Brehm, 1995; Gosling, Denizeau, & Oberlé, 2006). For example, if asked first about the importance of the question of tuition raises, they will say that it is of little importance (even though in questionnaires administered a few weeks previously they rated it as of high importance), thereafter going on to express an unchanged degree of opposition to the change and rating themselves as highly responsible for what they did.

The best explanation of these patterns of result is that subjects' mindreading systems automatically appraise them as having freely chosen to do something bad, resulting in negative affect. Then when confronted with the attitude questionnaire they rehearse various possible responses, responding affectively to each in the manner outlined by Damasio (1994), Gilbert & Wilson (2007), and others. They select the one that "feels right" in the circumstances, which is one that provides an appraisal of their actions as being significantly *less* bad. As a result of making that selection, their bad feelings go away. For example, saying (and hearing themselves say) that they do not oppose a raise in tuition (contrary to what they believe) enables their earlier actions to be appraised as *not bad*, and as a result they cease to feel bad. In contrast, it seems quite unlikely that subjects should really be changing their minds prior to selecting an answer on the questionnaire, with their novel belief then being available to be authoritatively reported. For we know for sure that they do not change their beliefs unless offered the chance to express them, and there is no plausible mechanism via which a question about one's beliefs should lead to the formation of a new belief in these circumstances (which can then be veridically reported).

Such results are deeply problematic for traditional accounts of self-knowledge. For one would think that a direct question about one's beliefs (e.g. about the goodness or badness of a tuition raise, or about the importance of the issue) would have the effect of activating the relevant belief from memory. There seems no reason why a judgment of this sort should remain unconscious or be otherwise inaccessible to the subject. However, if subjects had authoritative access to this activated belief, then it would be mysterious how they could at the same time express an inconsistent belief and make themselves feel better by doing so. For if they say one thing while being aware that they think something else, then they would be aware of themselves as lying. And that ought to make them feel worse, not better.

Dual method theories

Someone wishing to defend a traditional account of self-knowledge might acknowledge the soundness of the data on confabulation, while pointing out that they only show that people sometimes attribute thoughts to themselves on the basis of self-directed mindreading. Consistently with the data, one can maintain that sometimes people have direct access to their thoughts, whereas on other occasions they rely on self-directed mindreading. Views of just this sort are defended by Nichols & Stich (2003) and Goldman (2006). Plainly, more needs to be said. For an account that simply asserts that sometimes we rely on self-directed mindreading and sometimes on introspection makes no predictions about when confabulation might be expected to occur, and it therefore cannot explain the patterning of the confabulation data. What is needed is some specification of the circumstances in which each of the two methods will be employed.

Nichols & Stich (2003) draw a distinction between detecting one's mental states and explaining one's mental states (or one's behavior). Introspection can only do the former. This is because explanation presupposes causation, and yet the causal relations that obtain among our thoughts, and between our thoughts and our actions, surely cannot be introspected. In contrast, while mindreading cannot directly detect a mental state, explanation falls squarely within its domain. What Nichols and Stich propose, then, is that subjects will resort to self-directed mindreading whenever they are asked why they did something or why they think something. In such circumstances confabulation will occur whenever the cues available for mindreading are misleading ones. In contrast, when asked simply to report on a current or recent thought, they should be able to access it directly, and confabulation will *not* occur.

The distinction between detecting and explaining might well be capable of accommodating some of the data on confabulation. But it plainly cannot capture it all. In particular, it cannot account for any of the examples discussed above. For in the self-perception and dissonance studies subjects are just asked to say how strongly they believe something. No explanations are required. It therefore remains a mystery why subjects should opt to mindread themselves when (by hypothesis) their thoughts are directly available for report. It might be felt, however, that the dual-control studies of Wegner & Wheatley (1999) are different. For in this case subjects are asked to divide responsibility for an outcome between themselves and another agent, which requires a judgment of their respective causal contributions. Even so, it remains puzzling that people's judgments of causality should anchor so closely around the midpoint in the subject-controlled trials. For if their intentions were accessible to them through introspection, as traditional accounts suppose, then the temporal contiguity between these and the outcome should have given subjects a powerful sense of control.

Goldman (2006) does not address the problem of explaining the patterning in the confabulation data. But Goldman (2009) opts to say that introspection is employed for conscious thoughts, whereas self-directed mindreading is needed for unconscious ones. All instances of confabulation are therefore explained as occurring in circumstances where the relevant thoughts are not conscious. There are, though, two broad kinds of account of conscious thought, and each makes Goldman's reply problematic. One claims that conscious thoughts are thoughts that we know ourselves to possess, either in general, or in the right sort of direct non-interpretive way. The other claims that conscious thoughts are ones that are "globally broadcast" to a wide range of executive, affective, and inferential systems. Let us consider these in turn.

If conscious thoughts are ones that we know ourselves to possess in some manner or other (whether by introspection or via self-directed mindreading), then it will be of no help to appeal to

the conscious–unconscious distinction in explaining instances of confabulation. For if conscious thoughts can involve self-directed mindreading then we should expect confabulation to occur in these cases too. On the other hand, if conscious thoughts are ones that we know ourselves to possess directly and non-interpretively, then we are no closer to saying in what circumstances confabulation can be expected. For it is already agreed that confabulation results from self-interpretation. So to say that confabulation may be expected in connection with unconscious thoughts is just to say that self-interpretation may produce errors in cases where we rely on self-interpretation. This is, of course, circular.

The remaining option for Goldman is to say that we have direct access to globally broadcast thoughts, whereas we need to rely on self-directed mindreading for the remainder. One problem for this option is that there is little evidence that thoughts (judgments, decisions, and the rest) are ever globally broadcast in the way that sensory or sensory-involving states are. For all of the evidence that we have of global broadcasting in the brain pertains to sensory states (Baars, 1988, 1997, 2002, 2003; Dehaene & Naccache, 2001; Dehaene, Naccache, Cohen, Bihan, Mangin, Poline, et al., 2001; Dehaene, Sergent, & Changeux, 2003; Dehaene, Changeux, Naccache, Sackur, & Sergent 2006; Baars, Ramsoy, & Laureys, 2003; Kreiman, Fried, & Koch, 2003). Another problem is that the best-validated models that we have of working memory, which also seems to employ a global broadcasting architecture, assume that it always implicates the maintenance, rehearsal, and manipulation of sensory-involving representations, including visual imagery and inner speech (Baddeley, 2006; Müller and Knight, 2006; Postle, 2006; D’Esposito, 2007; Jonides, Lewis, Nee, Lustig, Berman, & Moore, 2008). Moreover, it would be problematic, in any case, for Goldman to explain why the subjects’ real thoughts in the confabulation experiments sketched in “Confabulation data” should *not* have been globally broadcast, if such a thing is possible at all. For everyone agrees that attention is the main determinant of global broadcast and entry into working memory, and in the circumstances of those experiments one would expect subjects to be attending to their judgments or intentions, since they were either asked directly about them, or knew that they would need to give a report just a few moments later.

I conclude that dual-method theories cannot account for the full extent of the confabulation data. As a result, the only theory that does so successfully is one that claims that self-directed mindreading is the only access that any of us *ever* has to our own thoughts. This account will be elaborated in Section 3, before some additional evidence is considered under “Dissociation data” and “Brain imaging evidence.”

The interpretive sensory-access account

According to the interpretive sensory-access theory sketched in Section 1 and developed in detail in Carruthers (2011), the mindreading system is arranged as one of the consumers of globally broadcast sensory-involving information in the brain. It evolved initially for other-directed social purposes, whether of a “Machiavellian” sort (Byrne & Whiten, 1988, 1997), or for purposes of cooperation and collaboration (Richerson & Boyd, 2005; Hrdy, 2009), or both. This requires it to have access to perceptual information about the world, although by default it would also have access to any form of globally broadcast representation (including the attended outputs of proprioception and other forms of bodily experience). As a result, the mindreading faculty will also have access to imagistic representations (whether visual, motor, or in inner speech or hearing), since these utilize the same mechanisms as perception and can be globally broadcast in the same way (Paulescu, Frith, & Frackowiak, 1993; Kosslyn, 1994; Shergill, Brammer, Fukuda, Bullmore, Amaro, Murray, et al., 2002; Kosslyn & Thompson, 2003). This means that attributions of sensory

states to oneself are comparatively direct and immediate, since such states are available to the mindreading faculty as input.

It is important to realize that the mindreading system will have access to more than just strictly sensory non-conceptual states. This is because conceptual information of varying degrees of abstractness is generally bound into the content of any given sensory state and broadcast along with it. Thus, Kosslyn (1994), for example, characterizes the early stages of visual processing as a continual “questioning” of non-conceptual visual input by conceptual systems, which seek a “best match” with their representations of what objects and events of the relevant kind should look like. When a match is found, it is bound into the content of the visual percept to be broadcast along with it for yet other conceptual systems to consume and draw inferences from. In this way, there can be a cascade of increasingly abstract concepts bound into any given perceptual state, as successive conceptual systems receive the products of earlier systems’ work, and categorize the input accordingly (Barrett, 2005). As a result, one doesn’t just see textured surfaces and shapes, one sees *a face*; and one doesn’t just see a face, one sees *one’s mother*; and so on. Likewise for hearing: one doesn’t just hear a stream of phonemes, one hears someone *calling one’s name*, for example.

The work of the mindreading faculty, too, can be bound into the contents of globally broadcast perception or imagery. As a result, one doesn’t just see someone’s arm moving in the direction of a transparent object, one sees her as *reaching for a drink*; and one doesn’t just hear a stream of phonemes when someone talks, but one hears him as *wanting to know the way to the church*; and so on, and so forth. Likewise one’s own outer or inner speech can be heard as *judging that the church is straight ahead*. In either case the only access that this gives one to an underlying attitude is interpretive in character, depending on the combined work of the mindreading and language faculties. Yet, of course, an item of inner speech is not itself an attitude of any sort. So the event of hearing oneself as judging that the church is straight ahead is not itself an event of judging anything. Rather, at best, it expresses or is caused by such a judgment.

On the interpretive sensory-access account, then, while one generally has direct (non-interpretive) knowledge of one’s own sensory-involving states, the only access that one has to propositional attitudes of judging, deciding, intending, and so on (whether one’s own or someone else’s) is interpretive, mediated by some form of sensory or imagistic awareness. The interpretive sensory-access theory comports well with global broadcasting accounts of the architecture of human cognition, as well as with widely accepted theories of working memory. It is also directly supported by the extensive confabulation data discussed earlier, since self-attributions of mental states will be subject to just the same sorts of errors of interpretation as attributions of mental states to other people. In contrast, no form of direct-access theory of self-knowledge has any of these benefits.

The interpretive sensory-access account is also supported by a widespread agreement among psychologists who study human metacognition or “thinking about [one’s own] thinking” (including judgments of learning, feelings of knowing, and confidence judgments). This is that metacognitive judgments are inferential and cue-based, relying on a variety of sensorily-accessible cues (Reder, 1987; Metcalfe, Schwartz, & Joaquim, 1993; Koriat, 1995, 1997; Dunlosky & Metcalfe, 2009). People rely on such things as feelings of familiarity, or the swiftness with which an answer comes to mind, when judging whether they know something, or when judging their degree of confidence. There is nothing here to suggest that they have direct access to their underlying states of mind. Yet these findings are, of course, just what the interpretive sensory-access theory would predict.

All the evidence considered so far is strongly supportive of the interpretive sensory-access account. But it remains to consider some other evidence that might be thought, on the contrary, to support the distinctive and separate character of self-knowledge.

Dissociation data

One way of showing that the interpretive sensory-access account is incorrect would be to demonstrate dissociations in one's competence to acquire self-knowledge and other-knowledge. The account predicts that these should not occur, since each form of knowledge is held to employ the same mindreading faculty utilizing the same sensory channels (albeit sometimes relying on different forms of evidence, such as inner speech or visual imagery in the case of self-knowledge). Just such claims of dissociation have been made by Nichols & Stich (2003), Goldman (2006), and Robbins (2009) in respect of either schizophrenia, autism, or both. The present section will discuss these syndromes in turn.

Schizophrenia

There is now extensive evidence of mindreading deficits in schizophrenia generally (see Brüne, 2005, and Sprong, Schothorst, Vos, Hox, & Van Engeland, 2007, for wide-ranging reviews of the existing literature). Indeed, even first-degree relatives of people with schizophrenia show mindreading deficits that are independent of age, education, and IQ (Janssen, Krabbendam, Jolles, & van Os, 2003). So one might wonder whether people with schizophrenia *also* show deficits in self-knowing. If they do not, as Robbins (2009) speculates, then this would present an anomaly for the interpretive sensory-access account.

A test of this hypothesis is provided by Koren, Seidman, Poyurovsky, Goldsmith, Viksman, Zichel, et al. (2004), Koren, Seidman, Goldsmith, & Harvey (2006), who used the Wisconsin Card Sorting Task (WCST) in conjunction with measures of metacognitive ability. Following each sorting of a card (and before receiving feedback), patients were asked to indicate their confidence in the correctness of their performance on a 100-point scale, after which they had to indicate whether they wanted that trial to count toward their final score (which would impact how much money they would win). Koren and colleagues looked especially for correlations between the various measures of performance and other measures that are known to be predictive of real-world competence and successful independent living. (Specifically, they used measures of insight into one's own illness and measures of competence to consent to treatment.) They found only small-to-moderate correlations between the basic WCST scores and the latter. However, the results from the measures of metacognitive ability correlated quite highly with the measures of successful real-world functioning. These findings have since been confirmed by Stratta, Daneluzzo, Riccardi, Bustini, & Rossi (2009). And in a separate experimental paradigm, Lysaker, Dimaggio, Carcione, Procacci, Buck, Davis, et al. (2010) found that measures of metacognitive self-awareness are a good predictor of successful work performance of people with schizophrenia over a 6-month period.

It would seem, then, that self-directed metacognitive abilities are inversely related to the severity of schizophrenic illness. This allows us to conclude that metacognitive abilities are generally damaged in people with schizophrenia; for the severity of their disease correlates with an increased inability to monitor their current mental lives and to choose adaptively as a result. This is just what would be predicted if both self-knowledge and other-knowledge utilize the same mindreading faculty, as the interpretive sensory-access theory suggests.

Nichols & Stich (2003) claim that a specific form of schizophrenia—namely, passivity schizophrenia—demonstrates a dissociation in the reverse direction. They think that these patients exhibit a failure of self-knowledge together with normal mindreading abilities. The first part of this claim has at least a superficial plausibility. For such people complain that their own actions aren't under their control. A patient might say, for example, "When I decide to comb my hair, it isn't me who controls the movement of my arm, but the FBI." Such patients are also apt to complain of

“hearing voices” (in reality their own self-generated inner speech), and they may believe that other people are inserting thoughts into their heads against their will.

There are two things wrong with Nichols and Stich's suggestion, however. One is that there is no reason to think that people with passivity schizophrenia have normal mindreading abilities. In part this criticism is motivated by the very strong association between schizophrenia and mindreading deficits generally, as discussed above. But it is also supported by an fMRI study conducted by Brüne, Lissek, Fuchs, Witthaus, Peters, Nicolas, et al. (2008), specifically with patients suffering from passivity kinds of schizophrenic illness. While these people succeeded on the simple mindreading tasks they were asked to complete, they employed a very different network of brain regions when doing so than do normal controls. This suggests that their mindreading *system* isn't normal, even if they are partly able to compensate in other ways.

In the second place, however, classic passivity symptoms are not best explained by the failure of a self-knowledge system. Rather they are better explained by the failure of one of the main components of the action-control system (Frith, Blakemore, & Wolpert, 2000a, b). This is a comparator mechanism that is hypothesized to receive a so-called “forward model” of the expected sensory consequences of movement, created from the “efference copy” of the motor instructions for that movement, comparing this with the afferent sensory feedback from the movement itself, and enabling one to make swift on-line corrections as the movement unfolds (Wolpert & Kawato, 1998; Wolpert & Ghahramani, 2000; Jeannerod, 2006). We know that this system is damaged in passivity forms of schizophrenia specifically. For patients with passivity symptoms are unable to make online corrections in their own movements in the absence of visual feedback (Frith, 1992). There is reason to think that systematic damage to the comparator system would give rise to experiences of the sort that might well issue in a sense of alien control, as I shall now explain.

One of the normal effects of the comparator system is to “damp down” conscious experience of any incoming perceptual information that matches the predictions of the forward model. This is because if everything is proceeding as expected then no attention needs to be paid to it. As a result, sensory experience of one's own movements is normally greatly attenuated. This is why it is impossible to tickle yourself (Blakemore, Frith, & Wolpert, 1998, 1999). It is also why someone unwrapping a candy at the theatre will barely hear the noise they are making, while those around them are greatly disturbed. It turns out, however, that patients with passivity forms of schizophrenia **can** tickle themselves, and their experiences of their own actions are not modulated by their motor intentions (Blakemore, Smith, Steel, Johnson, & Frith, 2000). Hence, they will experience their own movements with the same sort of sensory vividness as would be present if someone else were making their movements for them, and they will experience their own inner speech just as if another person were speaking. This is, of course, exactly what they report.

I conclude, therefore, that there is no reason to think that patients with schizophrenia (or specific forms of schizophrenia) demonstrate a dissociation between self-knowledge and other-knowledge. There is nothing, here, to challenge the interpretive sensory-access account.

Autism

Nichols and Stich (2003) and Goldman (2006) argue that autism represents a dissociation between mindreading (which is widely agreed to be damaged in this population) and self-awareness, which they claim remains intact. They place considerable reliance on a study by Farrant, Boucher, & Blades (1999), who tested children with autism (as well as learning-disabled and normal children matched for verbal mental age) on a range of metamemory tasks. Since they were able to find no significant differences between the groups, the authors conclude that metacognition is unimpaired

in autism. It should be emphasized, however, that almost all of the children with autism who participated in this study were sufficiently well advanced to be able to pass first-level false-belief tasks. So we should predict that they would have some understanding of their own minds, too, and that they should be capable of completing simple metacognitive tasks.

Moreover, none of the experimental tasks employed by Farrant and colleagues required subjects to attribute current or recently past thoughts to themselves. On the contrary, the tasks could be solved by anyone who possessed the requisite mental concepts who was also a smart behaviorist. For example, one experiment tested whether the children with autism were aware that it is easier to learn a small number of items than a larger number. Not surprisingly, the children did well on this test. For they would have had ample opportunity over a number of years of schooling to have established a reliable correlation between the number of items studied in a task and the number of responses that are later evaluated as correct. (Note that the average age of the children with autism in this experiment was eleven years.)

In contrast with the claims of Nichols & Stich (2003) and Goldman (2006), many studies have found paired deficits of mindreading and self-knowledge among children with autism. Some of these have looked at children's awareness of their own intentions. Thus, Williams & Happé (2010) used the knee-jerk response, for example, asking groups of children whether or not they had **meant** to move their leg. The children with autism were much worse than the control groups in identifying their knee-jerk as unintended, and in all groups success was highly correlated with success in a set of third-person false-belief tasks.

In a separate set of experiments, Williams & Happé (2010) measured capacities to attribute intentions in the third-person as well as in the first. Subjects were asked to complete a picture, such as a drawing of a girl with a missing ear, or a cup with a missing handle. But in each case they drew on a sheet of transparent acetate that had been laid over another, so that although they thought they were completing one picture, they were in fact completing a different one. For example, in drawing what they intended to be the ear on the side of a girl's head they had in fact drawn a handle on a cup. When the ruse was revealed to them, they were asked what they had meant to draw. They then watched a video of the same task being undertaken by another child, and were asked the same question in the third person.

The results of this experiment were that the children with autism were significantly worse at identifying both their own and others' intentions than were the ability-matched children with developmental delay. In both groups success was strongly correlated with success in a number of false-belief tasks. It would appear from these data that the capacity to attribute intentions to oneself is just as damaged in children with autism as is the capacity to attribute intentions to other people, and that both result from the difficulties that such children have with mindreading in general.

Other studies have looked at the capacity to attribute false beliefs to oneself and to others, often using the unexpected contents test (or "Smarties task"). Typically-developing children begin to pass both versions of this task at about the same age, normally around four (Wellman, Cross, & Watson, 2001). A number of experimenters have found that children with autism are equivalently delayed on this task for both *self* and *other* (Baron-Cohen, 1991, 1992; Russell & Hill, 2001; Fisher, Happé, & Dunn, 2005). Some, however, have found that performance is significantly **better** on the *self* question than on the *other* question, suggesting that self-awareness might be comparatively spared in autism (Perner, Frith, Leslie, & Leekam, 1989; Leslie & Thaiss, 1992).

Williams & Happé (2009) reasoned that the differentially better performance on the *self* question found in some studies might be due to the fact that the children are asked at the outset to **say** what they think is in the container. Children with autism might then succeed in the task by remembering what they had previously said, rather than by recalling or reasoning about their earlier belief.

Williams and Happé therefore devised a version of the task that would elicit belief spontaneously, without requiring any verbal expression. The experimenter pretended at the outset of the interview to have cut her finger, and asked the subject to fetch her a band aid, in circumstances where a number of different types of container were in plain sight, but out of the experimenter's reach. When the child opened the band-aid box, however, he would find that it contained crayons. The same *self* and *other* questions were then asked as usual. The results were that children with autism performed poorly in both versions of this task relative to controls.

In fact, Williams & Happé (2009) found that the children with autism experienced significantly more difficulty in the *self* version of the task than when predicting what another person would think. A similar finding is reported by Lombardo, Barnes, Wheelwright, & Baron-Cohen (2007). Their subjects with autism had significantly more impairment in measures of understanding their own emotions than they displayed with regard to other people's emotions. These findings might be thought to suggest a partial dissociation between self-knowledge and other-knowledge. A more plausible suggestion, however, is made by Williams and Happé. This is that whatever rules and heuristics the children with autism have learned in order to help them cope, and to enable them to attribute mental states to people, will generally be outward-looking in character and focused on the social world. For it is the social world that they find especially threatening and unpredictable. So the difference may be one of performance, and does nothing to suggest that competence in mindreading can be spared relative to competence in self-attribution.

Finally, it is worth mentioning some studies by Klein, Chan, & Loftus (1999), Klein, Cosmides, Costabile, & Mei (2002), Klein, Cosmides, Murray, & Tooby (2004) of an individual with autism, which are claimed to demonstrate a dissociation between self-knowledge and other-knowledge. Although this individual has severely impaired episodic recall, and fails to distinguish among the personalities of close family members, he has a stable model of his own personality traits that correlates pretty well with the estimates of those who know him best. It seems, then, that not only can reliable self-knowledge of traits be obtained in the absence of episodic memory, but also that it is independent of any capacity to gain knowledge of the personality traits of other people.

Knowledge of one's personality traits is not the same as knowledge of one's current or recently past thoughts, of course, which is our focus in this chapter. However, it might be thought to imply it. Knowing that people are acting selfishly, or generously, or stubbornly seems to require knowledge of their goals, as well as an understanding of their construal of the situation (their beliefs). So if our conceptions of people's personalities are built up gradually from our evaluations of their actions as they occur, then such knowledge would seem to presuppose a capacity to attribute current thoughts to the agents in question. It is not obvious, however, that one's beliefs about people's personalities are always constructed in this way, especially when that person is oneself. Rather, one's self-conception may initially be constructed, in whole or in part, from the evaluations of others. If one's parent comments, "Don't be so stubborn," this might lead one to encode, "I am stubborn." And once one has formed a stable self-conception, this will be apt to influence one's behavior in a self-fulfilling manner. Conceiving oneself to be stubborn, one will be apt to act stubbornly; believing oneself to be generous, one will be more likely to do generous things; and so on.

Thus, if the individual studied by Klien and colleagues had formed his self-conception in such a manner, then the degree of correlation with other's personality assessments of him can be explained without needing to suppose that he has the capacity to attribute current thoughts to himself at all. And we can also explain why his judgments of the personality traits of his family are comparatively undifferentiated. For it seems likely that children have many fewer opportunities to observe other peoples' personality-relevant evaluations of close family members than they are aware of receiving themselves.

Even if we suppose that the individual with autism studied by Klein et al. (1999, 2004) had developed his self-conception on the basis of piecemeal evaluations of his own actions, however, the discrepancy between his trait-knowledge for *self* and familiar *others* can be explained by the interpretive sensory-access account. Simplifying somewhat, in order to judge that other people are acting generously one needs to attribute to them knowledge that someone needs help, combined with sufficient motivation to provide that help despite significant costs to themselves. This will require mindreading. But in order to judge that one is oneself acting generously it is far from clear that one needs to attribute to oneself knowledge that someone needs help. Rather, the first-order fact that someone does need help will suffice, thereby reflecting one's knowledge without requiring one to metarepresent one's state of knowledge. And in order to know that one is overcoming a significant cost to oneself while providing that help one can rely on subjectively experienced feelings of affective conflict. These are, of course, only accessible to the mindreading system in the first person (consistently with the interpretive sensory-access account of self-knowledge). The discrepancy between this individual's knowledge of his own personality traits and the traits of his family members may therefore result from a difference in performance, not reflecting any difference in competence in attributing mental states within the two domains.

I conclude, therefore, that there is no reason to think that people with autism demonstrate a dissociation between self-knowledge and other-knowledge, any more than people with schizophrenia do.

Brain imaging evidence

A widespread consensus has emerged concerning the network of brain regions that is specifically implicated in third-person mindreading. These include the medial prefrontal cortex, posterior cingulate cortex, superior temporal sulcus, and temporo-parietal junction (Frith & Frith, 2003; Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Saxe, 2009; Lombardo, Chakrabarti, Bullmore, Wheelwright, Sadek, Suckling, et al., 2010). The question for us is whether the same, or a distinct, brain network is implicated in self-knowledge. We first consider studies that have paired *self* and *other* mental-state attribution tasks, before examining studies of metacognition.

Self and other

There have been remarkably few studies that have directly targeted our question. There have, however, been numerous imaging experiments of knowledge of personality traits in oneself and others (e.g. Kelley, Macrae, Wyland, Caglar, Inati, & Heatherton, 2002; Kjaer, Nowak, & Lou, 2002; Lou, Lubner, Crupain, Keenan, Nowak, & Kjaer, 2004; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004; Pfeifer, Lieberman, & Dapretto, 2007; but see Gillihan & Farah, 2005, for a powerful critique of the assumptions made by such studies). These are of little direct relevance for us, since no one thinks that a personality trait is the sort of thing that one can directly introspect. Even if the initial acquisition of trait-knowledge requires thought-attribution, the adults in these studies are likely to have well-established models of their own personality traits, in which case they can answer questions about themselves directly from memory without needing to reason at all (Klein & Lax, 2010). It is small wonder, then, that many studies find different patterns of activation in the two conditions—albeit with very little consistency across experiments.

In one of the very few studies to contrast third-person mindreading with attribution of current mental states to oneself, Ochsner, Knierim, Ludlow, Hanelin, Ramachandran, Glover, et al. (2004) scanned subjects while they viewed a series of photographs, in three separate conditions. In one, they had to judge their own emotional reaction to the image (pleasant, unpleasant, or neutral). In another, they had to judge the emotional reaction of a character depicted within the image

(pleasant, unpleasant, or neutral). And in the third base-line condition they had to judge whether the photograph had been taken indoors or outdoors. Many of the regions of the mindreading network were found to be active in common between the *self* and *other* conditions. These included medial prefrontal cortex, posterior cingulate, and the superior temporal sulcus.

However, *self* judgments activated medial prefrontal cortex to a greater extent than did *other* judgments. This effect is likely to result from the fact that medial prefrontal cortex seems to be active whenever one processes social information generally (Saxe & Powell, 2006), and because one would expect deeper and more elaborated processing in relation to the self (Gillihan & Farah, 2005). *Other* judgments, in contrast, distinctively activated an area of left lateral prefrontal cortex, which the experimenters interpret as an area implicated in maintaining and manipulating information about the external world. *Other* judgments also differentially activated an area of visual cortex, which the experimenters interpret as resulting from the greater attention paid to visual stimuli when judging the emotional state of another person. So there is nothing in these findings to suggest the existence of distinctive mechanisms for self-knowledge.

Most other studies that purport to contrast *self* and *other* mental-state attribution have failed to pair other-directed mindreading tasks with attributions of current mental states to oneself. For example, Saxe, Moran, Sholz, & Gabrieli (2006) claim to find areas of both overlap and non-overlap for *self* and *other*. However, the design of their study is an odd one. The *other* conditions are intended to test for false-belief reasoning. Subjects were scanned while reading either a false-belief story or a story involving a false photograph or map. As one might expect, the main elements of the mindreading network were active in this condition, including medial prefrontal cortex and the temporo-parietal junction bilaterally. In the *self* condition, in contrast, subjects read a series of trait adjectives, and either had to judge whether or not the adjective applied to themselves, or whether it was positive or negative. Since this task doesn't require one to attribute any current mental states to oneself, people will either answer from memory (using a stable self-model), or by mindreading and generalizing from items in episodic memory.

Likewise, Lombardo et al. (2010) conducted an extensive imaging study with a self-other design. In each case mentalizing judgments were contrasted with physical judgments. In the *self* condition, subjects had to use a four-point scale to answer questions like, "How likely are you to think that keeping a diary is important?" This was contrasted with physical questions like, "How likely are you to sneeze when a cat is nearby?" The *other* condition was identical, except that the questions all related to the Queen. (This study was conducted in the UK.) Note, however, that subjects weren't asked to make judgments about their current thoughts and attitudes. Rather, they were asked to estimate what their attitudes would be toward various suggested possibilities (such as keeping a diary). Since these questions might be ones that some subjects had never previously considered, they might have had to engage in the same sort of simulative reasoning process that they would use when trying to determine the likely attitudes of another person. Moreover, other subjects might have been able to answer the *self* questions directly from memory (for example, if they knew that they update a diary every day).

I conclude that while very few studies have contrasted the brain regions involved in mindreading with those that are active when one attributes a current or very recently past mental state to oneself, what evidence there is supports the interpretive sensory-access account.

Metacognition in the brain

Although many investigations of metacognition in the brain have failed to find activity in the mindreading network, this is likely to be an artifact of the experimental designs that have been

used. For instance, Maril, Simons, Weaver, & Schacter (2005) set out to differentiate between feelings of knowing and tip-of-the-tongue states. Since these are both metacognitive in nature, the interpretive sensory-access theory predicts that the contribution made by the mindreading system should be washed out when either one is subtracted from the other. Even when the brain activations involved in both of these kinds of feeling were combined together by the experimenters, they were contrasted with the combined “know” and “don’t know” responses. Of course these, too, are equally metacognitive. Likewise in the studies by Reggev, Zuckerman, & Maril (2011), when episodic and semantic feelings of knowing were combined together they were contrasted with the brain activity involved in the “don’t know” response. Since both sets of conditions involve metacognitive states, the interpretive sensory-access account predicts that activity should not be seen in the mindreading network.

Quite different results can be obtained when metacognitive judgments are contrasted with first-order ones. For example, Chua, Schacter, Rand-Giovannetti, & Sperling (2006) investigated the brain regions that are active when subjects make metacognitive confidence judgments. They contrasted judgments of confidence with first-order judgments of recognition. One form of differential activity was found in orbitofrontal cortex. While this lies outside the mindreading network, it nevertheless makes good sense. For this is one of the main brain regions where affective feelings are represented, and judgments of confidence are often grounded in feelings of confidence. But in addition, differential activity was found in posterior cingulate cortex and in regions of medial and lateral parietal cortex that include the temporo-parietal junction. Although the authors themselves don’t notice the point, these are vital elements of the mindreading network, as we noted earlier.

In a later study, Chua, Schacter, & Sperling (2009) contrasted metamemory judgments with two distinct kinds of first-order judgment, one of which consisted of judgments of recognition, as before, but the other of which involved judgments of facial attractiveness (which was used as an additional control). The investigators found differential activity in a number of areas. These included posterior cingulate and areas of medial and lateral parietal cortex that contain the temporo-parietal junction. But in addition they found activity in medial prefrontal cortex, which is also generally thought to be part of the mindreading network—albeit a region whose functions may also be somewhat more general. Almost all components of the mindreading network were thereby found to be active.

These results provide further support for the interpretive sensory-access theory, while being correspondingly problematic for those who believe that self-awareness is direct and independent of mindreading.

Conclusion

This chapter has contrasted two views of knowledge of one’s own thoughts. According to the first, self-knowledge of at least a subset of thoughts is direct, non-interpretive, and especially reliable. According to the second, self-knowledge results from turning our mindreading capacities on ourselves, utilizing sensory-involving cues (including visual imagery and inner speech as well as perceptions of our own behavior). These cues need to be interpreted, just as the mindreading system needs to interpret sensory input when attributing thoughts to other people. We have seen that the interpretive sensory-access account comports well with global broadcasting theories of the architecture of cognition, as well as with sensory-involving theories of working memory, and that it can explain the widespread data on confabulation for thoughts collected by social psychologists. In contrast, a direct-access account cannot explain this data. Moreover, there is no convincing evidence of dissociations between self-knowledge and other-knowledge in either

schizophrenia or autism, and nor do there appear to be different brain networks implicated in the two forms of knowledge. So the interpretive sensory-access theory is currently better supported by the evidence.

Acknowledgements

Some of the material in this chapter is drawn from Carruthers (2011), with permission of the author and Oxford University Press.

References

- Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.
- Baars, B. (1997). *In the Theatre of Consciousness*. Oxford: Oxford University Press.
- Baars, B. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences* 6:47–52.
- Baars, B. (2003). How brain reveals mind: Neuroimaging supports the central role of conscious experience. *Journal of Consciousness Studies* 10:100–14.
- Baars, B., Ramsoy, T., & Laureys, S. (2003). Brain, consciousness, and the observing self. *Trends in Neurosciences* 26:671–5.
- Baddeley, A. (2006). *Working Memory, Thought, and Action*. Oxford: Oxford University Press.
- Baron-Cohen, S. (1991). The development of theory of mind in autism: Deviance and delay. *Psychiatric Clinics of North America* 14:33–51.
- Baron-Cohen, S. (1992). Out of sight or out of mind: Another look at deception in autism. *Journal of Child Psychology and Psychiatry* 33:1141–55.
- Barrett, H. (2005). Enzymatic computation and cognitive modularity. *Mind and Language* 20:259–87.
- Bem, D. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review* 74:183–200.
- Blakemore, S.-J., Frith, C., & Wolpert, D. (1999). Spatiotemporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience* 11:551–9.
- Blakemore, S.-J., Smith, J., Steel, R., Johnson, E., & Frith, C. (2000). The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: Evidence for a breakdown in self-monitoring. *Psychological Medicine* 30:1131–9.
- Blakemore, S.-J., Wolpert, D., & Frith, C. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience* 1:635–40.
- Bolyard, C. (2009). Medieval skepticism. In: E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. <<http://plato.stanford.edu/archives/spr2009/entries/skepticism-medieval>>.
- Briñol, P., & Petty, R. (2003). Overt head movements and persuasion: a self-validation analysis. *Journal of Personality and Social Psychology* 84:1123–39.
- Brüne, M. (2005). “Theory of mind” in schizophrenia: A review of the literature. *Schizophrenia Bulletin* 31:21–42.
- Brüne, M., Lissek, S., Fuchs, N., Witthaus, H., Peters, S., Nicolas, V., Juckel, G., & Tegenthoff, M. (2008). An fMRI study of theory of mind in schizophrenic patients with “passivity” symptoms. *Neuropsychologia* 46:1992–2001.
- Byrne, R. & Whiten, A. (Eds) (1988). *Machiavellian Intelligence*. Oxford: Oxford University Press.
- Byrne, R. & Whiten, A. (Eds) (1997). *Machiavellian Intelligence II*. Cambridge: Cambridge University Press.
- Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.
- Caston, V. (2002). Aristotle on consciousness. *Mind* 111:751–815.

- Chua, E., Schacter, D., & Sperling, R. (2009). Neural correlates of metamemory: A comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of Cognitive Neuroscience* 21:1751–65.
- Chua, E., Schacter, D., Rand-Giovannetti, E., & Sperling, R. (2006). Understanding metamemory: Neural correlates of the cognitive process and subjective level of confidence in recognition memory. *NeuroImage* 29:1150–60.
- D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B* 362:761–72.
- Damasio, A. (1994). *Descartes' Error*. London: Papermac.
- Dehaene, S., & Naccache, L. (2001). Toward a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79:1–37.
- Dehaene, S., Changeux, J-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences* 10:204–11.
- Dehaene, S., Naccache, L., Cohen, L., Bihan, D., Mangin, J., Poline, J., & Riviere, D. (2001). Cerebral mechanisms of word priming and unconscious repetition masking. *Nature Neuroscience* 4:752–8.
- Dehaene, S., Sergent, C., & Changeux, J. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences* 100:8520–5.
- Descartes, R. (1641). *Meditations on First Philosophy*. In E. Anscombe & P. Geach (Eds & transl.), *Descartes Philosophical Writings*. London: Thomas Nelson & Sons (1954).
- Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. London: Sage Publications.
- Elliot, A., & Devine, P. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology* 67:382–94.
- Farrant, A., Boucher, J., & Blades, M. (1999). Metamemory in children with autism. *Child Development* 70:107–31.
- Fisher, N., Happé, F., & Dunn, J. (2005). The relationship between vocabulary, grammar, and false belief task performance in children with autistic spectrum disorders and children with moderate learning difficulties. *Journal of Child Psychology and Psychiatry* 46:409–19.
- Frith, C. (1992). *The Cognitive Neuropsychology of Schizophrenia*. Hillsdale: Erlbaum.
- Frith, C., Blakemore, S-J., & Wolpert, D. (2000a). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews* 31:357–63.
- Frith, C., Blakemore, S-J., & Wolpert, D. (2000b). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London B* 355:1771–88.
- Frith, U., & Frith, C. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358, 459–73.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences* 2:493–501.
- Gilbert, D., & Wilson, T. (2007). Prospection: Experiencing the future. *Science* 317:1351–4.
- Gillihan, S., & Farah, M. (2005). Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin* 131:76–97.
- Goldman, A. (2006). *Simulating Minds*. Oxford: Oxford University Press.
- Goldman, A. (2009). Replies to the commentators. *Philosophical Studies* 144:477–91.
- Gosling, P., Denizeau, M., & Oberlé, D. (2006). Denial of responsibility: A new mode of dissonance reduction. *Journal of Personality and Social Psychology* 90:722–33.
- Hrdy, S. (2009). *Mothers and Others*. Cambridge: Harvard University Press.
- Janssen, I., Krabbendam, L., Jolles, J., & van Os, J. (2003). Alterations in theory of mind in patients with schizophrenia and nonpsychotic relatives. *Acta Psychiatrica Scandinavica* 108:110–17.
- Jeannerod, M. (2006). *Motor Cognition*. Oxford: Oxford University Press.

- Jonides, J., Lewis, R., Nee, D., Lustig, C., Berman, M., & Moore, K. (2008). The mind and brain of short-term memory. *Annual Review of Psychology* 59:193–224.
- Kelley, W., Macrae, C., Wyland, C., Caglar, S., Inati, S., & Heatherton, T. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience* 14:785–94.
- Kjaer, T., Nowak, M., & Lou, H. (2002). Reflective self-awareness and conscious states: PET evidence for a common midline parietofrontal core. *NeuroImage* 17:1080–6.
- Klein, S., Chan, R., & Loftus, J. (1999). Independence of episodic and semantic self-knowledge: The case from autism. *Social Cognition* 17:413–36.
- Klein, S., Cosmides, L., Costabile, K., & Mei, L. (2002). Is there something special about the self? A neuropsychological case study. *Journal of Research in Personality* 36:490–506.
- Klein, S., Cosmides, L., Murray, E., & Tooby, J. (2004). On the acquisition of knowledge about personality traits: Does learning about the self engage different mechanisms than learning about others? *Social Cognition* 22:367–90.
- Klein, S., & Lax, M. (2010). The unanticipated resilience of trait self-knowledge in the face of neural damage. *Memory* 18:918–48.
- Koren, D., Seidman, L., Goldsmith, M., & Harvey, P. (2006). Real-world cognitive—and metacognitive—dysfunction in schizophrenia: A new approach for measuring (and remediating) more “right stuff.” *Schizophrenia Bulletin* 32:310–26.
- Koren, D., Seidman, L., Poyurovsky, M., Goldsmith, M., Viksman, P., Zichel, S., & Klein, E. (2004). The neuropsychological basis of insight in first-episode schizophrenia: A pilot metacognitive study. *Schizophrenia Research* 70:195–202.
- Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General* 124:311–33.
- Koriat, A. (1997). Monitoring one’s own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General* 126:349–70.
- Kosslyn, S. (1994). *Image and Brain*. Cambridge: MIT Press.
- Kosslyn, S., & Thompson, W. (2003). When is early visual cortex activated during visual mental imagery. *Psychological Bulletin* 129:723–46.
- Kreiman, G., Fried, I., & Koch, C. (2003). Single neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Sciences* 99:8378–83.
- Leslie, A., & Thaiss, L. (1992). Domain specificity in conceptual development: Evidence from autism. *Cognition* 43:225–51.
- Locke, J. (1690). *An Essay Concerning Human Understanding*, J. Yolton (Ed.). London: J. Dent & Sons (1965, two volumes). .
- Lombardo, M., Barnes, J., Wheelwright, S., & Baron-Cohen, S. (2007). Self-referential cognition and empathy in autism. *PLoSOne* 9:e833.
- Lombardo, M., Chakrabarti, B., Bullmore, E., Wheelwright, S., Sadek, S., Suckling, J., MRC AIMS Consortium, & Baron-Cohen, S. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience* 22:1623–35.
- Lou, H., Luber, B., Crupain, M., Keenan, J., Nowak, M., & Kjaer, T. (2004). Parietal cortex and representation of the mental self. *Proceedings of the National Academy of Sciences USA* 101:6827–32.
- Lysaker, P., Dimaggio, G., Carcione, A., Procacci, M., Buck, K., Davis, L., & Nicolo, G. (2010). Metacognition and schizophrenia: The capacity for self-reflectivity as a predictor for prospective assessments of work performance over six months. *Schizophrenia Research* 122:124–30.
- Macrae, C., Moran, J., Heatherton, T., Banfield, J., & Kelley, W. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex* 14, 647–54.
- Maril, A., Simons, J., Weaver, J., & Schacter, D. (2005). Graded recall success: An event-related fMRI comparison of tip of the tongue and feeling of knowing. *NeuroImage* 24:1130–8.

- McCloskey, D., Colebatch, J., Potter, E., & Burke, D. (1983). Judgments about onset of rapid voluntary movements in man. *Journal of Neurophysiology* 49:851–63.
- Meltzoff, A., & Brooks, R. (2008). Self-experience as a mechanism for learning about others. *Developmental Psychology* 44:1257–65.
- Mendelson, M. (2009). Saint Augustine. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Available at: <<http://plato.stanford.edu/archives/fall2009/entries/augustine/>>.
- Metcalf, J., Schwartz, B., & Joaquim, S. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19:851–61.
- Müller, N., & Knight, R. (2006). The functional neuroanatomy of working memory: contributions of human brain lesion studies. *Neuroscience* 139:51–8.
- Nichols, S., & Stich, S. (2003). *Mindreading*. Oxford: Oxford University Press.
- Ochsner, K., Knierim, K., Ludlow, D., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S. (2004). Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience* 16:1746–72.
- Paulescu, E., Frith, D., & Frackowiak, R. (1993). The neural correlates of the verbal component of working memory. *Nature* 362:342–5.
- Perner, J., Frith, U., Leslie, A., & Leekam, S. (1989). Explorations of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development* 60: 689–700.
- Pfeifer, J., Lieberman, M., & Dapretto, M. (2007). “I know you are but what am I?”: Neural bases of self- and social knowledge retrieval in children and adults. *Journal of Cognitive Neuroscience* 19:1323–37.
- Postle, B. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience* 139:23–38.
- Reder, L. (1987). Strategy selection in question answering. *Cognitive Psychology* 19:90–138.
- Reggev, N., Zuckerman, M., & Maril, A. (2011). Are all judgments created equal? An fMRI study of semantic and episodic metamemory predictions. *Neuropsychologia* 49:3036–45.
- Richerson, P., & Boyd, R. (2005). *Not By Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.
- Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience* 2:661–70.
- Robbins, P. (2009). Guilt by dissociation: Why mindreading may not be prior to metacognition after all. *Behavioral and Brain Sciences* 32:159–60.
- Russell, J., & Hill, E. (2001). Action-monitoring and intention reporting in children with autism. *Journal of Child Psychology and Psychiatry* 42:317–28.
- Saxe, R. (2009). Theory of mind (neural basis). In: W. Banks (Ed.), *Encyclopedia of Consciousness*, Vol. 2, 401–10 Cambridge: MIT Press.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind.” *NeuroImage* 19:1835–42.
- Saxe, R., & Powell, L. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science* 17:692–9.
- Saxe, R., Moran, J., Sholz, J., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Scan* 1:229–34.
- Scher, S., & Cooper, J. (1989). Motivational basis of dissonance: The singular role of behavioral consequences. *Journal of Personality and Social Psychology* 56:899–906.
- Shergill, S., Brammer, M., Fukuda, R., Bullmore, E., Amaro, E., Murray, R., & McGuire, P. (2002). Modulation of activity in temporal cortex during generation of inner speech. *Human Brain Mapping* 16:219–27.
- Simon, L., Greenberg, J., & Brehm, J. (1995). Trivialization: The forgotten mode of dissonance reduction. *Journal of Personality and Social Psychology* 68:247–60.

- Sprong, M., Schothorst, P., Vos, E., Hox, J., & Van Engeland, H. (2007). Theory of mind in schizophrenia: Meta-analysis. *British Journal of Psychiatry* 191:5–13.
- Stratta, P., Daneluzzo, E., Riccardi, I., Bustini, M., & Rossi, A. (2009). Metacognitive ability and social functioning are related in persons with schizophrenic disorder. *Schizophrenia Research* 108:301–2.
- Wegner, D., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of the will. *American Psychologist* 54:480–91.
- Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72:655–84.
- Wells, G., & Petty, R. (1980). The effects of overt head movements on persuasion. *Basic and Applied Social Psychology* 1:219–30.
- Williams, D., & Happé, F. (2009). “What did I say?” vs. “What did I think?”: Attributing false beliefs to self amongst children with and without autism. *Journal of Autism and Developmental Disorders* 39:865–73.
- Williams, D., & Happé, F. (2010). Representing intentions in self and other: Studies of autism and typical development. *Developmental Science* 13:307–19.
- Wolpert, D., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience* 3:1212–17.
- Wolpert, D., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks* 11:1317–29.
- Young, M. (1995). On the origin of personal causal theories. *Psychonomic Bulletin and Review* 2:83–104.

This page intentionally left blank

Index

Note: page numbers in *italics* refer to figures and tables. Footnotes are indicated by the suffix 'n' followed by the note number, for example 448n1.

A

- ABAT* 331
- abstraction, infants 24
- accidents, moral judgements 94–6
- acquisition question 448
- action chaining, studies in autism 389
- action discrimination, mirror neurons 237
- action prediction 97, 98, 99–100, 136
- action rationality studies, autism 390, 392
- action sequence studies, autism 388
- action understanding
- in autism
 - behavioral studies 387–8
 - broken mirror theory 386–7
 - implicit measures 388–9
 - mentalizing theory 385–6, 387
 - neuroimaging studies 389–90, 391
- brain and cognitive systems 384
- infants 246–7
- non-mirror mechanisms 276
- role of mentalizing 382–4
- role of mirror neurons 249, 381–2, 384
- humans 275–6
 - monkeys 267–70, 272–3
 - role of superior temporal sulcus 148
- activational effects on development 309
- affective content, functional imaging studies 152
- affective empathy 326
- see also* emotional empathy
- affiliation 97–8, 99–100
- AGRN* 331
- alexithymia 203, 206
- alpha activity, resting state EEG studies 126–7
- amniocentesis studies of hormone levels 310–11
- limitations 314–15
- amygdala 182
- effect of testosterone administration 317
 - and emotional empathy 181–2, 368–70
 - and empathy for fear 455
 - integrated emotion systems model 365, 366, 372
 - role in stress response 294
 - and social anxiety disorder 300
 - Wolframin expression 334–5
- androgens
- effects on females 308
 - clinical conditions 310
 - see also* testosterone
- androstenedione, levels in autism 333
- anger, empathy for 202
- anger recognition 365
- effect of testosterone administration 316
- anterior cingulate cortex (ACC) 182
- involvement in empathy 198, 201
 - involvement in pain perception 181, 200
 - see also* cingulate cortex
- anterior insula (AI)
- involvement in empathy 198, 201–4
 - involvement in pain perception 200
 - see also* insula
- anterior midcingulate cortex (aMCC)
- involvement in empathy 198, 202, 202–4
 - involvement in pain perception 200
 - see also* cingulate cortex
- anthropomorphism 98
- anticipation, role of mirror neurons 244
- anticipatory intervening paradigm 5, 7
- antisocial personality disorder (ASPD) 365
- appearance–reality (A–R) tasks, teaching interventions
- in autism 417
- apraxia, action recognition problems 249
- AQ-Child (Autistic Spectrum Quotient-Child Version)
- scores, relationship to fetal testosterone levels 313, 314
- AR* 332
- ARNT2* 331, 334
- Asperger's syndrome
- candidate gene association study 329–35
 - emotion understanding, teaching interventions 419
 - empathy for pain 406
 - implicit mentalizing 385
 - moral judgements 96
 - see also* autism
- attachment
- in autism 400
 - role of oxytocin 296
- autism 413
- action understanding
 - behavioral studies 387–8
 - implicit measures 388–9
 - neuroimaging studies 389–90, 391
 - communication and language 404–5
 - non-literal language understanding 64
 - dissociation 475–8
 - effect of oxytocin administration 299
 - emotion recognition impairment 182
 - empathy 206, 367
 - empathy profile 326
 - false belief understanding 59
 - functional imaging studies 139, 184–5, 406
 - future research directions 392

- autism (*cont'd*)
- gaze direction perception 123–4
 - genetic studies 327–9
 - candidate genes 331–2
 - endophenotypes 335–6
 - gene association study design 329–30, 332
 - study results 332–5
 - imitation, controlled studies 402
 - mirror neuron system function 185, 253–4, 281–3, 405–6
 - relationship to fetal testosterone levels 313, 314
 - self, development of 407–8
 - self-awareness
 - clinical descriptions 398
 - first-hand accounts 399
 - of mental states 403–4
 - self-conscious emotions, controlled studies 400–2
 - self–other relations, controlled studies 399–400
 - theories of 330
 - broken mirror theory 386–7
 - mentalizing theory 385–6, 387
 - theory of mind interventions 414, 422
 - benefits 423
 - developmental approach 417–19
 - teaching false belief understanding 414–17
 - teaching joint attention 420–2
 - teaching social skills 420
 - automatic mindreading 78–9, 80, 173
 - cognitive efficiency 81–2
 - limitations 82, 164
 - unnecessary or unhelpful processes 80–1
 - aversive conditioning 368–9
 - taste aversion learning 370–1
 - AVPR1 genes 331
 - AVPR1A 335
 - AVPR1B 335
 - awareness of events
 - infants' understanding of 53–4
 - see also* perspective-taking
- B**
- Bayesian learning 20–1, 453–4
 - domains 28
 - framework principles 21, 22
 - understanding of personality traits 27–9
 - BDNF 331
 - behavior-reading, infants 7–10
 - belief
 - direct and indirect tests 41n5
 - dissociation 41–2
 - ERP studies 119–22
 - Introduction and Elimination rules 37–9
 - role in moral judgments 94–5
 - supposition versus simulation 39–40
 - see also* false belief understanding
 - belief-desire psychology 51
 - belief-desire tasks, adult studies 75, 76–7
 - role of working memory 77–8
 - Bem Sex Role Inventory score, influence of fetal testosterone levels 312
 - biases
 - adults' judgements 74
 - see also* egocentric bias
 - blindfold studies of infants 9–10, 22–6
 - gaze-following 53
 - blindsight 41n5
 - bluff, functional imaging studies 139
 - body-directed encoding, VIP neurons 270, 272
 - bonding, role of oxytocin 296
 - borderline personality disorder (BPD), effect of oxytocin 300
 - brain
 - cortical organization 146–7
 - functional heterogeneity of neurons 145–6
 - localization of function 144–5
 - neural basis of mindreading 86–8
 - see also* functional imaging studies
 - scale of neural responses 145
 - brain–behavior correlations 223–5
 - brain lesion studies 164–5
 - amygdala 181–2, 455
 - of emotional empathy 181
 - frontal lobe 178
 - insula and basal ganglia 454
 - left temporoparietal lesion 167–9
 - right lateral prefrontal cortex lesion 166–7
 - semantic dementia 171
 - working memory impairment 169–70
 - see also* case studies; stroke damage; transcranial magnetic stimulation
 - broadly congruent mirror neurons 235–6, 238, 243
 - Broca's aphasia 249
 - broken mirror theory of autism 386–7, 390–1
- C**
- Cambridge Child Development Project 311–14
 - canonical neurons 267
 - case studies
 - CM (semantic dementia) 171, 172
 - Phineas Gage 178
 - NK (insula damage) 454
 - NM (amygdala damage) 455
 - PF (left temporo-parietal lesion) 167–9, 170
 - SM (bilateral amygdala damage) 182, 455
 - WBA (right lateral prefrontal brain lesion) 166–7
 - CAST (Childhood Autism Spectrum Test) score, relationship to fetal testosterone levels 313, 314
 - castration experiments 308
 - caudate, integrated emotion systems model 366
 - causal explanation
 - children's' understanding 44–6
 - role of belief, Introduction and Elimination rules 37–9
 - teleology 35–7
 - cerebellum, mirror neurons 241
 - cerebrospinal fluid (CSF), oxytocin levels 292
 - CGA 332
 - CGRP 332
 - charity, principle of 37
 - children
 - belief 40–1
 - developmental pattern 84–6
 - dissociation 41–2
 - false belief understanding 54–6, 133
 - early sensitivity 42–4
 - ERP studies 121–2

- implicit understanding 354–6
 - functional imaging studies, developmental differences 142
 - late childhood development 63–4, 65
 - mirror neuron system activity 246
 - probabilistic models of learning 20–1
 - resting state EEG studies 127–8
 - teleology 35–7, 47
 - understanding of caused behavior 44–5
 - understanding of personality traits 26–9
 - understanding of reasons for action 45–6
 - see also* deaf children; infants
 - “child-scientist” theory-theory 450–1, 453
 - chimpanzees
 - cultural profiles 438
 - imitation 439–40, 441
 - mindreading 434–6, 441–2
 - China, theory of mind development sequence 57–8
 - cingulate cortex
 - activation during metacognition 480
 - involvement in empathy 198, 200, 201–2, 204
 - involvement in pain perception 181, 200
 - in psychopathy 367
 - classical test theory (CTT) 109
 - CNR1* 331, 335–6
 - cognitive empathy 180, 182–4, 326, 366
 - interaction with emotional empathy 187–9, 188
 - in psychopathy 366–7
 - collaborative behaviour
 - effect of testosterone 317
 - functional imaging studies 99
 - hunter-gatherers 432
 - common primate ancestors 438–41
 - communication in autism 404–5
 - non-literal language understanding 64
 - teaching interventions 418
 - Communication of Affect Receiving Ability Test (CARAT) 105, 106, 114
 - comparator system 475
 - compassion, relationship to empathy 196
 - competitive situations, functional imaging studies 98, 99
 - compulsive imitation 253
 - concern, empathic 207
 - conduct disorder 365
 - confabulation 468–70
 - dual-method theories 471–2
 - conformity 437
 - congenital adrenal hyperplasia (CAH) 310, 334
 - connectivity patterns, cortical 147
 - contextual factors, influence on mindreading 79
 - contextual information, modulation of empathic responses 206, 218
 - cooperation
 - effect of testosterone 317
 - functional imaging studies 99
 - hunter-gatherers 432
 - copy number variations (CNV), in autism 328
 - core knowledge theories 19
 - cortical organization 146–7
 - cortisol response, effect of oxytocin 293
 - coyness, in autism 401
 - CU (callous and unemotional) traits, psychopathy 364, 365
 - cue-based paradigms, empathy for pain 198
 - computational route 200–1
 - cultural differences
 - in executive-function development 56
 - in false belief understanding 55
 - in social attribution 29
 - in theory of mind development sequence 57–8, 65
 - cultural transmission 436–7
 - culture 434
 - hunter-gatherers 433
 - reconstructing ancestral cultural capacities 438–41
 - “curse of knowledge” effect 74
 - CYP* genes 332
 - CYP11A* 334
 - CYP11B1* 333, 333–4
 - CYP17A1* 333
- ## D
- deaf adults
 - functional imaging studies 139
 - late acquisition of false belief reasoning 352
 - deaf children 345, 358
 - background variables 346, 347
 - educational approaches 346
 - executive function 351–2
 - development 355, 356
 - false belief understanding 19
 - implicit understanding 354–6
 - study results 349–50
 - imitation 353
 - joint attention limitations, DoH children 356–8
 - non-literal language understanding 64
 - research methods 348
 - seeing-knowing understanding 354
 - theory of mind development 59–61, 65–6
 - pretense understanding 62–3
 - understanding of desires and intentions 353–4
 - variation in language experience 346, 347, 348–9
 - deception understanding
 - in frontotemporal dementia 187
 - functional imaging studies 99
 - decoupling
 - functional imaging studies 141
 - infants 12
 - “deep social mind”
 - culture 436–7
 - evolution, role of hunter-gatherer existence 431–4
 - default network 461–2
 - deictic terms, usage in autism 404–5
 - dementia
 - frontotemporal 187
 - semantic 171
 - depression, impaired emotion recognition 182
 - desire reasoning, deaf children 353–4
 - development 54–6, 55, 84–6, 127–8
 - activational effects 309
 - in autism 407–8
 - cultural differences 56, 57–8, 65
 - deaf children 59–61, 62–3, 65–6, 355, 356
 - effects of hormones 309, 311–314

- cultural differences (*cont'd*)
 of executive function 56, 351–2, 355, 356
 functional imaging studies 142
 late 63–4, 65
 of mirror neuron system 245–7
 probabilistic models 21–2
 puberty 315–15, 319
 two-system account 83–6
DHCR7 331
 Diagnostic Analysis of Non-Verbal Accuracy Scale
 (DANVA, DANVA-2-AF) 105, 106
 reliability 108, 109, 110, 114
 “difference-making” approach, causal explanation 37
 diffusion imaging (DTI) 147
 digit length ratios 316
 disgust
 empathy for 202, 280–1, 454
 in psychopathy 370–1, 372
 integrated emotion systems model 365, 366
 dissociation 41–2, 47, 474
 in autism 475–8
 between mirroring and self-projection 462
 in schizophrenia 474–5
 dissonance studies 469–70
 distance, relationship to mirror neuron responses 270
 distress, empathic 207
 diverse beliefs (DB) awareness 57, 58
 diverse desires (DD) awareness 57
 domains 28
 dorsal anterior cingulate cortex (dACC)
 involvement in empathy 202–4
see also cingulate cortex
 dorsal premotor cortex, mirror activity 275
 dorsolateral prefrontal cortex (dlPFC) 183, 184
 Down syndrome, theory of mind interventions 416
 dual-method accounts of self-knowledge 468, 471–2
- E**
 eating, imagination study 457–8
 EEG studies *see* electroencephalography
 effortful processes 76–80, 164
 role of TPJ 153
 egalitarianism 434
 egocentric bias 74, 77, 452–3
 after stroke damage 88
 electrode implantation studies, mirror neuron system
 241
 electroencephalography (EEG) 3, 119
 resting state EEG studies
 adults 126–7
 children 127–8
 mu-suppression 128–9
 studies of mirror neuron system 240, 241, 242
 in autism 253, 281–2
 children 246
see also event-related potential (ERP) studies;
 mu-suppression
 Elimination rules 37–9
 embedded beliefs
 brain lesion studies 169–70
 functional imaging studies 154
 embedded motor actions, mirror neuron response
 237–8
- embodied judgmental accuracy 111
 embodiment theories 215
 emotional contagion 179
 relationship to empathy 195–6
 emotional empathy 179, 180–2, 248, 366
 functions 368
 interaction with cognitive empathy 187–9, 188
 in psychopathy
 anger processing 371
 disgust processing 370–1
 fear and sadness processing 368–70
 pain processing 370
see also empathy
 emotional experience, role of insula 202, 203
 emotional intelligence 105n2
 emotional responses, role of ACC and AI 204
 emotion recognition
 coactivation of ESS and MSAS 221
 ERP studies 124
 functional imaging studies 137, 138, 139, 150–1
 integrated emotion systems model 365–6
 interventions in autism 419
 role of mirror neurons 280–1
 role of oxytocin 297–8
 emotions, self-conscious, studies in autism 400–2
 empathic concern 207
 empathic distress 207
 empathy 178–80, 214, 366, 437
 behavioral consequences 207
 brain-behavior correlations 223–5
 cognitive 182–4
 components of 326
 correlation with mirror neuron activity 244, 247
 definitions 195–7
 emotional 180–2
 ESS and MSAS systems 215–217
 coactivation 221
 dissociation 217–20
 functional coupling 222
 functional imaging studies 198–200, 199, 202–5
 different computational routes 200–1
 distinctions between direct and vicarious pain
 201–2
 future research areas 207–8, 225–6
 heritability 327
 candidate genes 331–2
 endophenotypes 335–6
 gene association study design 329–30, 332
 insights from autism genetics 327–9
 study results 332–5
 influence of fetal testosterone levels 312–13, 314
 interaction between emotional and cognitive systems
 187–9, 188
 modulation of responses 205–6
 neuropsychiatric defects 184–7
see also autism; psychopathy; schizophrenia
 and oxytocin 298
 picture-based and cue-based paradigms 197–8
 self-conscious emotions, studies in autism 400–2
 shared network hypothesis 197, 198
see also cognitive empathy; emotional empathy;
 motor empathy
 empathy quotient (EQ) 329

- emulation 250
 - non-human primates 438–40
- EN-2(AUTS1)* 331
- enactment imagination (E-imagination) 456–8
- encoding of experience 13–14
- environment monitoring 165, 174
 - brain lesion study 167–9
- envy, effect of oxytocin 296
- equivalent dipole problem
 - EEG studies 128
 - ERP studies 124
- ESR* genes 332
 - ESR2* 333
- event-related potential (ERP) studies 88
 - advantages 125–6
 - of belief/desire reasoning 119–21
 - in children 121–2
 - of mental state decoding 122–4
 - technical problems 124–5
- evolutionary theories, “mesh” with simulation theory 462–3
- evolution of the mind
 - reconstructing ancestral cultural capacities 438–41
 - role of hunter-gatherer existence 431–4
- executive function
 - in autism, teaching interventions 415–16
 - development 56, 351
 - deaf children 351–2, 355, 356
 - in psychopathy 365
 - role in mindreading 77–8
- experience sharing 215–17
- experience sharing system (ESS) 216, 218, 225
 - coactivation with MSAS 221
 - dissociation from MSAS system 217–20
 - functional coupling with MSAS 222
- eye closure, infants’ understanding 22
- Eyes *see* Reading the Mind in the Eyes test
- eye tracking studies
 - in Asperger’s syndrome 385
 - in autism 388–9
 - deaf children 356
 - infants 5
- F**
- F5 266–7
 - classes of neurons 235–6
 - kinematic processing 382
 - mirror neurons 233–4
 - see also* mirror neuron system
- facial expression perception 180, 249
 - anger, in psychopathy 371
 - in autism 254
 - disgust 202, 280–1, 454
 - in psychopathy 370–1
 - effect of oxytocin 297–8
 - effect of testosterone administration 316
 - fear and sadness processing 368–9
 - impairment in psychopathy 369–70
 - modulation of empathic responses 205
 - see also* emotion recognition
- false-belief tasks
 - factors in success 12–14
 - infants’ failure in 10–12
 - unexpected events 148
- false belief understanding 57, 168, 179
 - adult studies 75, 76–7
 - role of working memory 77–8
 - in autism 59, 385, 403, 476–7
 - developmental approach 417–19
 - teaching interventions 414–17
 - brain lesion studies 170
 - left temporoparietal lesion 167–9
 - right lateral prefrontal cortex lesion 166–7
 - chimpanzees 436
 - cognitive versus affective 183
 - cultural differences in age of achievement 55
 - deaf children 19, 59
 - late acquisition 352
 - methodological issues 345–9
 - role of executive functioning 351–2
 - study results 349–50
 - direct and indirect tests 41n5
 - early sensitivity 42–4
 - ERP studies 119–22
 - functional imaging studies 133–4, 136
 - study design 140
 - implicit understanding 354–6
 - infants 4–7, 14–15, 54–6, 133
 - approaches 44–6
 - blindfold studies 9–10
 - ignorance rule 7
 - Mistaken Max 35–6
 - and moral judgements 94–5, 96
 - non-automatic nature 78–9
 - role of executive functioning 351–2, 351–2
 - see also* belief
- familiarization-test paradigms 52
- fear, empathy for 202, 365, 368–9, 455
 - impairment in psychopathy 369–70
- fear recognition, role of amygdala 182
- fetal androgen theory 330
- fetal testosterone levels 309, 310–11
 - Cambridge Child Development Project 311–14
 - limitations of amniotic fluid measurements 314–15
- “fight-or-flight” response 293
- first-person mindreading 448n1
- fish, social learning 436–7
- framework principles 21, 22
- framework theories 20
- frontal lobe lesions 178, 183
 - compulsive imitation 253
- frontal lobes
 - and belief reasoning 120–1, 122
 - role in mindreading 88
 - see also* inferior frontal gyrus; orbitofrontal cortex; prefrontal cortex
- frontotemporal dementia 187
- FSHB* 332
- functional adaptation analysis 155
- functional imaging studies 88, 132, 143, 183
 - in autism 184–5, 254, 282, 389–90, 391, 406
 - of cooperation and competition 98, 99
 - differences between theory of mind regions 151–3
 - of empathy 180–1, 198–205, 199, 280–1

functional imaging studies (*cont'd*)
 brain-behavior correlations 223–5
 ESS and MSAS systems 215–22
 of false belief understanding 133–4
 inter-individual variability 142
 limitations 156
 magnitude of responses 153–4
 of mentalizing 383
 of mirror neuron system 239, 241–2, 274–6, 279, 381–2, 384
 in autism 254, 282
 of moral judgement-making 95, 96–7
 of pantomiming 222
 patterns within ToM regions 154–6
 sample stimuli 135, 136
 in schizophrenia 186
 of self-knowledge 478–80
 of social concepts 171
 of social inference 222
 strong hypothesis 143–4
 empirical objections 149–51
 theoretical objections 144–9
 study designs 134–40
 necessary and sufficient features 140
 of testosterone administration 317
 use of transcranial magnetic stimulation 143

G

GABR genes 331
GABRB3 334
 Gage, Phineas 178
 gaze direction, role of amygdala 369
 gaze direction perception, ERP studies 123–4
 gaze-following, infants 53
 gender-typical behavior, influence of fetal testosterone levels 312, 314
 generosity, effect of oxytocin 296
 genome wide association studies (GWAS), of autism 328
 goal processing 382
 grammar, developmental associations with mindreading 83–4
 grammatical processing defects, false belief reasoning 170
 gray matter volume, ventromedial prefrontal cortex 186
 guilt, in autism 401

H

habituation-test paradigms (familiarization-test paradigms) 52
 heartbeat detection task 202
 Hebbian learning 245–6, 251
 helping behavior
 children 42–3
 role of empathy 207
 hidden emotion (HE) awareness 57
 hierarchical Bayesian modelling (HBM) 21
 high-level mindreading 172
 simulation theory process 451–4
 theory-theory process 450–1
 hormones 308
 effects on development 309

future research areas 318–19, 320
 postnatal effects, studies of activational hormones 315–16
see also oxytocin; testosterone
 how, what and why of actions, mirror neuron responses
 humans 243
 monkeys 238
HOXA1 331, 334
HSD genes 332
 hunter-gatherers 431–3

I

IGF1, *IGF2* 331
 ignorance rule, false-belief tasks 7
 imagination
 memory distortion 458n5
 power of 456–8
 role in simulation theory process 452, 453
 imitation
 in autism 386, 387–8, 402
 compulsive 253
 deaf children 353
 non-human primates 438–40
 over-imitation 437–8, 440–1
 role in cultural transmission 437
 role of mirror neurons 249–50, 278–9
 imitation learning 279–80
 implicit mentalizing
 in autism 385
 and simulation theory 459–60
 impression management 100
 indirect elicited-response task 11–12
 infants
 action understanding 246–7
 blindfold studies 22–6
 evidence for mind reading 3–5, 14–15
 behavior-reading interpretations 7–10
 false-belief task failures 10–12
 predictive behavior 5–7
 false belief understanding 54
 early sensitivity 42–4
 gaze-following 53
 imitation 440
 intention understanding 51–3
 mindreading abilities 84–5
 mindreading system, alternative developmental accounts 84–6
 mirror neuron system development 245–7
 mu-suppression studies 128–9
 testosterone levels 309
 understanding of agents' awareness of events 53–4
 understanding of perception 22–6
 inferior frontal gyrus (IFG) 88, 182
 integrated emotion systems model 365, 366
 involvement in empathy 180–1, 200
 kinematic processing 381–2
 inferior parietal lobe (IPL) 180, 182, 265
 and belief reasoning 120, 122
 mirror neurons 267, 274–5, 381, 386
 insula 182, 280–1
 integrated emotion systems model 365, 366
 involvement in empathy 198, 201–4
 involvement in pain perception 181, 198–200, 201–2

involvement in taste aversion learning 370–1
 mirroring of disgust 454
 integrated emotion systems (IES) model 365–6, 372–3
 intention understanding 469
 in autism 403, 476
 children 41, 45–6, 47, 51–3
 deaf children 353
 role in moral judgments 94–5
 role of mirror neurons 272–3
 see also false belief understanding; Mistaken Max
 interpersonal closeness, functional imaging studies 152
 Interpersonal Perception Test 15 (IPT15) 105, 106, 114
 Interpersonal Perception Test 30 (IPT30) 105, 106, 114
 interpreters, use in studies of deaf children 348
 interpretive sensory-access theory 472–3
 intranasal oxytocin administration 292
 intraparietal sulcus
 goal processing 382
 involvement in empathy 200
 mirror neurons 234
 see also mirror neuron system
 Introduction rules 37–9
 Iran, theory of mind development sequence 58
 irony, children's understanding 63–4
 item response theory (IRT) 109, 114–15

J

jealousy, in autism 401–2
 joint actions, role of mirror neurons 244, 252
 joint attention
 limitations in deaf (DoH) children 356–8
 teaching interventions in autism 420–2, 423
 jokes, children's understanding 63–4
 judgmental accuracy 104
 item response theory (IRT) 114–15
 peer ratings versus self-ratings 104–5
 Social Relations Model (SRM) 115–16
 judgmental accuracy measures 105–6
 improvement of reliability 108
 “easy” tests 109–10
 embodied JA 111
 using psychometric correct answer 110–11
 reliability 114
 estimates of 106–8
 validity 108

K

kinematic processing 381–2
 “knee-jerk task” 403, 476
 knowledge, curse of 452–3
 see also egocentric bias
 knowledge access (KA) awareness 57, 58
 “knowledge task,” performance of deaf children 354

L

language 434
 in autism 404–5
 non-literal language understanding 64
 teaching interventions 418
 hunter-gatherers 433
 language development, relationship to fetal testosterone levels 311, 314

language evolution, role of mirror neurons 250–1
 “laser-beam points,” children with autism 405
 lateral intraparietal area (LIP) 266
 mirror neurons 267
 late slow wave (LSW) effect, ERP studies 122, 125, 126
 learning by observation, role of mirror neurons 252
 LHB 332
 LHCGR 332
 LHRHR 332
 “like-me” framework principle 22, 24–5
 limbic system, and emotional empathy 181–2
 localization of brain function 144–5
 long-term knowledge of ToM 165, 170–1, 170–2, 174
 looking in expectation, children 42
 looking-time studies 3, 4–5, 42
 interpretation 5
 low-level mindreading 173, 449, 454–6

M

magnetic resonance imaging (MRI)
 studies of mirror neuron system 239–40, 241–2, 243
 see also functional imaging studies
 magnetoencephalography (MEG)
 studies in autism 389
 studies of mirror neuron system 240, 241, 277–8
 MAOB 331, 334
 mastery, attainment over others 98
 medial prefrontal cortex (mPFC) 87, 98, 99, 128, 133, 183
 differences from other ToM regions 152–3, 155
 involvement in empathy 200, 217
 mentalizing 382–3
 role in moral judgment 96
 and self-knowledge 479, 480
 MEG studies *see* magnetoencephalography
 memory
 in autism 403–4
 long-term knowledge impairment 171–2
 role in mindreading 77–8
 social, effect of oxytocin 298
 working memory impairment 169–70
 mentalizing 276, 382–4, 462n12
 relationship to empathy 195
 relationship to fetal testosterone levels 311–12, 314
 role of mirror neurons 252–3
 mentalizing theory of autism 385–6, 387, 390–1
 functional imaging studies 390, 391, 406
 mental state attribution 217
 ERP studies 122–4
 functional imaging studies 155
 resting state EEG studies 126–7
 mental state attribution system (MSAS) 217, 218, 225
 coactivation with ESS 221
 dissociation from ESS system 217–20
 functional coupling with ESS 222
 mental state, self-awareness, autism 403–4
 metacognition 140n2
 functional imaging studies 480
 metaphor, children's understanding 63
 meta-representation, functional imaging studies 141

- mindreading 72
 - acquisition 458–61
 - cognitively efficient, but inflexible thinking 80
 - evidence for cognitive efficiency 81–2
 - limitations 82
 - unnecessary or unhelpful processes 80–1
 - criteria for an adequate theory 448
 - flexible and effortful thinking 75–7, 79–80
 - contextual and motivational influences 79
 - non-automatic nature 78–9
 - role of memory and executive function 77–8
 - functions 72–3
 - hunter-gatherers 433
 - in non-human primates 434–6, 441–2
 - simulation theory
 - high-level mindreading 451–4
 - low-level mindreading 454–6
 - power of imagination 456–8
 - study methods 74–5
 - trade-offs 73–4
 - two-system account 83
 - implications for developmental studies 83–6
 - implications for neural basis of mindreading 86–8
 - see also* action understanding; empathy; false belief understanding; self-knowledge
 - mindreading brain network 86–8
 - mirroring 128–9, 461, 462
 - simulation theory 454–6
 - “mirroring first” model 384
 - mirror neurons 267
 - mirror neuron system 19, 173, 180–1, 197, 216, 264–5, 380–2, 384, 437
 - broadly congruent neurons 235–6
 - connections with STS 236
 - dysfunctions
 - in autism 185, 253–4, 281–3, 386–7, 389–90, 391, 405–6
 - compulsive imitation 253
 - in schizophrenia 186
 - functions 248–9
 - action recognition 249
 - emotion recognition 280–1
 - empathy 244, 367–8
 - imitation 249–50, 278–80, 439
 - learning by observation 252
 - mentalizing 252–3
 - prediction 251–2
 - role in language evolution 250–1
 - humans 273–4
 - development 245–7
 - encephalographic studies 277–8
 - evidence for existence 238, 240–1
 - localization 241–2, 274–5
 - plasticity 245
 - properties 242–4
 - role in action understanding 275–6
 - single neuron studies 278
 - study techniques 239
 - TMS studies 276–7
 - location 234
 - monkeys 233–6
 - area F5 266–7
 - circuit for grasping 268
 - functional properties 267–73
 - organization of motor cortex 265–6
 - parietal areas and STS 267
 - properties 236–8
 - strictly congruent neurons 236
 - view dependence 270, 271
 - VIP neurons 270, 272
 - mirror-touch synaesthesia 247
 - Mistaken Max 35–6, 41, 45–6
 - supposition versus simulation 39–40
 - M-like states 452
 - modularity theories 19, 25, 451
 - modular processing
 - limitations 82
 - unnecessary or unhelpful processes 80–1
 - molecular genetics, study of oxytocin system 292
 - morality 93
 - impression management 100
 - morality on the ground 97–100
 - moral judgements 93–7
 - functional imaging studies 137, 142, 143, 152, 156
 - in psychopathy 371–3
 - motivation, influence on mindreading 79
 - motor cortex, organization in the monkey 265–6
 - motor empathy 366
 - and mirror neurons 367–8
 - in psychopathy 368
 - multi-voxel pattern analysis (MVPA) 155
 - mu-suppression 128–9, 181
 - studies in autism 185, 253, 281–2, 389, 406
 - studies of mirror neuron system 241, 242, 246, 253, 277–8, 281–2
- N**
- native signing deaf children
 - theory of mind development 59–60
 - see also* deaf children
 - neonatal imitation 246
 - nerve growth factor (NGF) 334
 - neural basis of mindreading 86–8
 - simulation theory 461–2
 - neural connectivity theory of autism 330
 - neural plausibility question 448
 - neural resonance 216
 - neurexins 334
 - neuroimaging *see* functional imaging studies
 - neurons, functional heterogeneity 145–6
 - NGF 331
 - NLGN genes 331
 - NLGN4X 334
 - non-congruent F5 neurons 233, 234, 235
 - non-human primates
 - imitation 438–9
 - mindreading 434–6, 441–2
 - see also* chimpanzees; orang-utans
 - non-literal language, children’s understanding 63–4
 - non-verbal communication, in autism 404
 - non-verbal stimuli, functional imaging studies 134, 136, 137, 138, 139
 - NRCAM 331
 - NTF3, NTF5 331
 - NTRK genes 331
 - NTRK1 333, 334

O

object-directed actions, mirror neuron response
 humans 242–3
 monkeys 236–7

object permanence 3

occlusion, mirror neuron response 237

OPRM1 331

orang-utans, imitation 439

orbitofrontal cortex (OFC) 183

 activation during metacognition 480

 effect of testosterone administration 317

 effects of brain lesions 186–7

 in psychopathy 367

organizational effects on development 309

orientation, correlation with search 8

“ought,” reasons versus rationality 36

outcomes, role in moral judgments 94–5

over-imitation 437–8, 440–1

OXT 331, 335

OXTR (oxytocin receptor gene) 331, 334, 335

polymorphism 292

 influence on attachment and bonding 296

 influence on emotion recognition 297

 influence on stress response 294

oxytocin (*OXT*) 335

 comparison with testosterone 317–18

 future research areas 302

 neurophysiology 291

 research methods 291–2

 and social approach 292–3

 role in attachment and social bonding 296

 role in interpersonal trust 294–6, 295

 stress-reducing effect 293–4

 and social cognition

 emotion recognition and empathy 297–8

 modulation of social memory 298

 therapeutic use 299–301

oxytocin system, integrative model 301

P

P3 component, ERP studies 126

pain

 congenital insensitivity to 248, 455

 mirror responses 278

pain, empathy for 181, 197, 222, 247–8, 454–5

 in Asperger's syndrome 406

 functional imaging studies 198–200, 199

 different computational routes 200–1

 distinctions between direct and vicarious pain 201–2

 involvement of anterior insula 202–4

 involvement of cingulate cortex 204

 modulation of responses 205–6

 picture-based and cue-based paradigms 197–8

 in psychopathy 370

 shared network hypothesis 198

pain matrix 198–200, 222

pain perception, role of insula 202

 pantomiming, functional imaging studies 222

 parallel forms reliability, JA measures 107–8

 parental care, role of testosterone 317

 parietal lobe 266

 activation during metacognition 480

 mirror activity 274–5

 mirror neurons 234, 236, 241, 242, 267

see also mirror neuron system

 projections to premotor cortex 265

see also inferior parietal lobe (IPL)

pars opercularis, involvement in empathy 200

passivity schizophrenia 474–5

pattern classification fMRI, studies of mirror neuron system 240, 242

“people search in the last place that they saw something” rule 8

perception, infants' understanding 22–6

perception-action hypothesis 180

perception-like nature of mindreading

 cognitive efficiency 81–2

 limitations 82

 unnecessary or unhelpful processes 80–1

personality traits 26

 children's understanding 26–9, 477

perspectives

 adults' understanding 77

 studies 80–1

 children's understanding 40, 46

 encoding of 13–14

 self-perspective inhibition, brain lesion study 166–7

 supposition 40

perspective-taking 437

 in autism, teaching interventions 418

 deaf children 354

 ERP studies 125–6

 functional imaging studies 150–1

 modulation of empathic responses 206, 207

 power of imagination 456–8

phenomenology 264

phobias, impaired emotion recognition 182

phrenology 144

picture-based paradigms, empathy for pain 197–8

 computational route 200–1

picture-in-the-head technique, teaching false belief understanding 415, 418

pity, relationship to empathy 196

plasticity, mirror neuron system 245

polycystic ovary syndrome (PCOS) 310, 333

POR 332

positron emission tomography (PET) *see* functional imaging studies

posterior cingulate 99

 involvement in empathy 217

see also cingulate cortex

posterior insula

 role in pain perception 198–200, 201

see also insula

posterior superior temporal cortex, involvement in empathy 200

precentral gyrus, mirror activity 274

precuneus (PC) 133

 differences from other ToM regions 152, 153

 involvement in empathy 200

 magnitude of responses 153–4

 mentalizing 382

- prediction
 - role of insula 203
 - role of mirror neurons 244, 251–2
 - predictive behavior, infants 5–7
 - prefrontal cortex 183–4
 - brain lesion study 166–7
 - functional studies in autism 184–5
 - see also* medial prefrontal cortex (mPFC)
 - premotor cortex
 - mirror neurons 233, 241, 242, 275, 381
 - see also* mirror neuron system
 - monkeys
 - area F5 266–7
 - organization 265–6
 - preschool development 54–6, 65
 - cultural variation 57–8
 - resting state EEG studies 127–8
 - sequence of understandings 57
 - see also* infants
 - pretend states, simulation theory 452, 453
 - pretense understanding, deaf children 62–3
 - pride, in autism 401
 - probabilistic models of development 21–2
 - probabilistic models of learning 20–1
 - understanding of personality traits 27–9
 - Profile of Non-verbal Sensitivity (PONS) 105, 106
 - reliability 108, 109, 110, 114
 - prosocial behaviors 224–5, 326
 - brain-behavior correlations 224–5
 - psychometric truth 110–11
 - psychopathy 95, 186–7, 206
 - cognitive empathy 366–7
 - emotional empathy
 - anger processing 371
 - disgust processing 370–1
 - fear and sadness processing 368–70
 - pain processing 370
 - empathy profile 326
 - integrated emotion systems model 365–6
 - moral judgement 371–3
 - motor empathy 368
 - neurocognitive impairments 364–5
 - psychophysics, study of mirror neuron system 239
 - pubertal development 315–16
 - future research areas 319
 - Punch and Judy, children's responses 12, 14
 - pyramidal tract neurons, mirror activity 270
- Q**
- Q-CHAT (Quantitative Checklist for Autism in Toddlers) score, relationship to fetal testosterone levels 313, 314
- R**
- RAPGEF4* 331
 - rationalizing 36
 - Reading the Mind in the Eyes test 105, 106, 106
 - effect of oxytocin 297
 - ERP studies 123
 - familiarity of performance 327
 - influence of fetal testosterone levels 311–12
 - influence of testosterone administration 316
 - reliability 114
 - resting state EEG studies 126–7
 - reality bias 74
 - referential expressions interpretation, children 42
 - repetition-suppression analysis 155
 - study of mirror neuron system 239–40, 242, 243
 - resonance theories 19
 - reward responses, role of testosterone 317
 - right temporo-parietal junction (RTPJ) 87, 99, 128, 133, 183, 184
 - activation during metacognition 480
 - activation in non-ToM tasks 149–50
 - differences from other ToM regions 152–3, 155
 - diffusion imaging (DTI) 147
 - functional imaging studies 139
 - in autism 406
 - involvement in empathy 200, 217
 - mentalizing 382–3
 - right, effect of fetal testosterone 312
 - role in moral judgment 95, 96, 96–7
 - “stimulus space” 147–9
 - lesion studies 143
 - risk prediction, role of insula 203
- S**
- sad events, functional imaging studies 152
 - sadness, empathy for 365, 368–9
 - impairment in psychopathy 369–70
 - “Sally and Anne” story 4, 7, 165–6, 179
 - Sam the Mouse 41
 - sarcasm understanding
 - children 63–4
 - in frontotemporal dementia 187
 - schizophrenia 336
 - dissociation 474–5
 - effect of oxytocin administration 300
 - impaired emotion recognition 182, 185–6
 - SCP2 332
 - seeing-knowing understanding, deaf children 354
 - selection processor 451
 - self, concept of 397
 - self-awareness 397
 - in autism
 - clinical descriptions 398
 - development 407–8
 - first-hand accounts 399
 - self-awareness of mental states 403–4
 - self-conscious emotions 400–2
 - self-experience, role in infants' understanding 24–6
 - self-knowledge 467–8, 480–1
 - confabulation theories 468–70
 - dissociation 474
 - in autism 475–8
 - in schizophrenia 474–5
 - dual-method theories 471–2
 - functional imaging studies 478
 - of metacognition 480
 - paired self and other tasks 478–9
 - interpretive sensory-access theory 472–3
 - self-other relations, in autism 399–400
 - self-perspective inhibition 165, 174
 - brain lesion study 166–7

- self-projection 461, 462
 - semantic dementia, case study 171
 - sensation-seeking, adolescents 316
 - sensorimotor field, insula 280
 - shared network hypothesis of empathy 197, 198
 - SHBG* 332
 - signature limits 86
 - sign language use
 - deaf children 346, 347, 348
 - relationship to theory of mind development 59–60
 - similarity to self, functional imaging studies 152
 - simulation theory 39, 179
 - acquisition of mindreading 458–61
 - bi-level approach 449
 - foundations 449
 - high-level mindreading 451–4
 - low-level mindreading 454–6
 - “mesh” with evolutionary theory 462–3
 - neural basis of mindreading 461–2
 - power of imagination 456–8
 - single neuron studies, mirror neuron system
 - in humans 278
 - in monkeys 266–7
 - SLC6A4* 335
 - SLC25A12*, *SLC25A13* 332
 - slow wave effect, ERP studies 120–1
 - smell perception, mirror properties 181
 - social anxiety disorder (SAD), effect of oxytocin
 - administration 299–300
 - social approach, role of oxytocin
 - attachment and social bonding 296
 - interpersonal trust 294–6, 295
 - stress-reducing effect 293–4
 - social attribution, understanding of personality traits 27–9
 - social behaviors, brain-behavior correlations 223–5
 - social cognition 366
 - role of oxytocin
 - emotion recognition and empathy 297–8
 - modulation of social memory 298
 - social-cognitive component of mindreading 449
 - social connection, role of mindreading 98–100
 - social development, relationship to fetal testosterone levels 311, 314
 - social-emotional responsivity theory of autism 330
 - social inference, functional imaging studies 222
 - social knowledge, association with temporal poles 171–2
 - social learning 439–41
 - fish 436–7
 - social-perceptual component of mindreading 449
 - Social Relations Model (SRM) 111, 115–16
 - social skills, teaching interventions in autism 420
 - social threat recognition, effect of testosterone
 - administration 316–17
 - socio-cognitive niche 441
 - cognitive and behavioural characteristics 432–4
 - origins in hunter-gatherer existence 431–2,
 - socio-economic backgrounds, deaf children 346, 347
 - somatosensory cortex
 - mirror neurons 241
 - role in empathy 204–5, 216
 - role in pain perception 198–200, 201
 - sounds, mirror neuron response
 - humans 243–4, 275–6
 - monkeys 237, 269
 - source-localization analysis, EEG studies 128
 - split-half reliability, JA measures 107–8
 - SRD5A1*, *SRD5A2* 332
 - SS1 and SS2 subsystems 43–4
 - Sternberg & Smith, tests of judgmental accuracy 105, 106, 114
 - “Sticker Test” 404
 - “stimulus spaces” 147–9
 - stress response, effect of oxytocin 293–4
 - strictly congruent mirror neurons 236, 238, 243
 - stroke damage
 - action recognition problems 249
 - functional imaging studies 143
 - left temporo-parietal lesion 167–9
 - mindreading impairment 88
 - right lateral prefrontal cortex lesion 166–7
 - STS* 332
 - subsequent memory paradigm 223
 - SULT2A1* 332
 - superior parietal lobe, mirror activity 275
 - superior temporal cortex, in psychopathy 368
 - superior temporal sulcus (STS) 133, 148, 183
 - connections with mirror neuron system 236, 267
 - supplementary motor cortex, mirror neurons 241
 - suppositional reasoning 39–40
 - supramarginal gyrus, involvement in empathy 200
 - sympathy, relationship to empathy 196
 - synaesthesia, mirror-touch 247
- T**
- TAC1* 331
 - tactical deception, non-human primates 434–5
 - tactile processing, mirror properties 181
 - task-execution question 448
 - taste aversion learning 370–1
 - teaching, role in cultural transmission 437
 - teleology 35–7, 47
 - teleology-in-perspective 40, 46
 - temporal poles 183
 - involvement in empathy 200, 217
 - mentalizing 382
 - role in social knowledge 171–2
 - temporo-parietal junction (TPJ) 87, 99, 133, 136
 - see also right temporo-parietal junction
 - testosterone
 - fetal and infant levels 309
 - future research areas 318–19, 320
 - postnatal effects 319
 - activational effects 316
 - administration studies 316–17
 - comparison with oxytocin 317–18
 - prenatal effects 319, 330
 - amniotic fluid measurements 310–11
 - Cambridge Child Development Project 311–14
 - limitations of amniotic fluid measurements 314–15
 - studies in clinical conditions 310
 - test-retest correlation, JA measures 107

- theory of mind (ToM) *see* mindreading
 - theory of mind components 165, 174
 - low- and high-level 172–3
 - Theory of Mind Scale sequence 57
 - cultural variation 57–8, 65
 - pretense understanding 63
 - progression of deaf children 60–2
 - theory-theory 19, 25, 179
 - “child-scientist” theory 450–1, 453
 - mindreading process 450
 - modularity theory 451
 - thinking, children’s understanding 63
 - thought bubbles, teaching false belief understanding 415
 - tool use, evolution 438
 - touch, empathy for 204–5, 247
 - traditions, reconstructing cultural capacities 438
 - training studies 19–20
 - blindfolds 22–6
 - mirror neuron system 245
 - trait bias 27, 28
 - transcranial magnetic stimulation (TMS)
 - of inferior frontal gyrus 381–2
 - to RTPJ 149
 - studies in autism 389, 406
 - studies of mirror neuron system 238, 239, 240, 249, 276–7
 - during imitation 279
 - studies of moral judgement 97, 143
 - TrkA 334
 - TRPV1 331
 - trusting behavior
 - effect of testosterone administration 317
 - role of oxytocin 294–6, 295, 318
 - TSPO 332
 - twins
 - studies of empathy 327
 - studies of hormonal effects 316
- U**
- unexpected events, functional imaging studies 147–8, 153
 - unusual actions, mentalizing 383
- V**
- VEGF 331
 - ventral intraparietal area (VIP), mirror activity 270, 272
 - ventromedial brain lesions 181
 - ventromedial prefrontal cortex (vmPFC) 183–4, 188
 - functional imaging studies, in autism 406
 - functional studies in autism 184–5
 - gray matter volume 186
 - integrated emotion systems model 366
 - in psychopathy 372, 373
 - VGF 331
 - video clips
 - mirror neuron responses, monkeys 270
 - theory of mind interventions in autism 416–17
 - view dependence, mirror neuron responses 270, 271
 - violation of expectation paradigm 42
 - VIPR1 331
 - virtual reality environments, use in functional imaging studies 136–7
 - visceral awareness 202
 - visual inference model 384
 - visual perception, mirror properties 181
 - visuo-motor priming 173
- W**
- WFS1 (Wolframin gene) 331, 334–5
 - Williams-Beuren syndrome 336
 - Wisconsin Card Sorting Task, study in schizophrenia 474
 - Wolframin 334–5
 - working memory
 - brain lesion studies 169–70
 - role in ToM 165, 174