JAMIE WARD

The Student's Guide to Social Neuroscience

The Student's Guide to Social Neuroscience

http://www.psypress.com/social-neuroscience-textbook/

The Student's Guide to Social Neuroscience

Jamie Ward

Psychology Press
Taylor & Francis Group
HOVE AND NEW YORK

http://www.psypress.com/social-neuroscience-textbook/

First published 2012
by Psychology Press
27 Church Road, Hove, East Sussex, BN3 2FA

Simultaneously published in the USA and Canada
by Routledge [or] Psychology Press
711 Third Avenue, New York, NY 10017

*Psychology Press is an imprint of the Taylor & Francis Group,
an Informa business*

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Ward, Jamie.
The student's guide to social neuroscience / Jamie Ward.
p. cm.
Includes bibliographical references and index.
ISBN 978–1–84872–004–6 (hardcover) – ISBN 978–1–84872–005–3 (pbk.)
1. Perception. 2. Social interaction. I. Title.
BF311.W26947 2012
153–dc22 2011009617

ISBN: 978–1–84872–004–6 (hbk)
ISBN: 978–1–84872–005–3 (pbk)

Typeset by Newgen Imaging Systems (P) Ltd, Chennai, India
Cover deisgn by Andy Ward

Printed and bound in the UK by

Contents

Preface

This textbook came about through a desire to create an accompanying text to *The Student's Guide to Cognitive Neuroscience* specifically in the area of social neuroscience. Cognitive neuroscience may be the parent discipline of social neuroscience, but it was becoming increasingly clear over the last few years that social neuroscience had now grown up and was trying to establish a home of its own. For example, there are now several excellent journals dedicated to it and many universities have introduced social neuroscience onto the undergraduate curriculum as a separate module distinct from cognitive neuroscience. This textbook aims to reflect the new maturity of this discipline and it attempts to convey the excitement of this field to undergraduate and early-stage postgraduate students.

My own interest in the field stemmed from the claims surrounding mirror systems, empathy, and theory of mind. At the start of this project, I imagined that this would form the core of the textbook. However, the more that I delved into the literature, the more I was taken aback by the volume and quality of research in other areas such as prejudice, morality, culture, and neuro-economics. The resulting book is, I hope, a more balanced view of the field than I initially anticipated. As with my previous textbook, it is not an exhaustive summary of the field. It is not my aim to teach students everything about social neuroscience but it is my aim to provide the intellectual foundations to acquire that knowledge, should they wish to become researchers themselves. My ethos is to try to present the key findings in the field, to develop critical thinking skills, and to instill enthusiasm for the subject.

In the absence of previous textbooks on social neuroscience, it was an interesting exercise deciding how to carve the field into chapters, and how to order the chapters. For example, 'Relationships' (Chapter 8) appeared and disappeared several times, being divided amongst 'Interactions' (Chapter 7) and 'Development' (Chapter 11). The first two chapters begin with an overview of the topic (Chapter 1) and a summary of the methods used in social neuroscience (Chapter 2). The 'methods' chapter is a condensed, but updated, version of the more extensive chapters in *The Student's Guide to Cognitive Neuroscience* and uses examples from the social neuroscience literature to illustrate the various methods. The third chapter covers the evolution of social intelligence and culture, and introduces mirror neurons in the context of imitation, social learning, and tool use. The fourth and fifth chapters deal with the 'primitive' building blocks of social processes, namely emotions and motivation (Chapter 4), and recognizing others (Chapter 5). Chapter 6 is concerned with empathy, theory of mind, and autism. The next two chapters consider social interactions (Chapter 7) and relationships (Chapter 8), dealing with issues such as altruism, game theory, attachment, and social exclusion. Chapter 9 is concerned with groups and identity, covering the notion of 'the self', prejudice, and religion. Chapter 10 covers antisocial behavior, aggression, and morality. The final chapter considers social development from infancy through to adolescence.

It will be interesting to see how the field of social neuroscience changes in the coming years. What new chapters will be added to subsequent editions of the book? Which chapters will require revising the most?

Finally, I would like to thank the many reviewers who provided constructive feedback on drafts of the chapters, and also Psychology Press for being so accommodating.

Jamie Ward

Brighton, UK, January 2011

CHAPTER 6

CONTENTS

Understanding others

6

If you see someone yawning do you yawn too? Most people probably do to some extent. Some behavior, such as laughing and yawning, is socially contagious. But can any wider significance be attached to such findings? One study of contagious yawning in chimpanzees speculates that 'contagious yawning in chimpanzees provides further evidence that these apes possess advanced self-awareness and empathic abilities' (Anderson, Myowa-Yamakoshi, & Matsuzawa, 2004). Another study, this time on humans, administered tests requiring reasoning about the mental states of other people (e.g. beliefs, knowledge) as well as measuring yawn contagion, and concluded that 'contagious yawning may be associated with empathic aspects of mental state attribution' (Platek, Critton, Myers, & Gallup, 2003). Of course, there is unlikely to be anything special about yawning itself. There might be a general tendency to *simulate* the behavior of others on ourselves (internally in our minds and brains) even if we do not overtly *reproduce* it (as observable behavior on our bodies). Thus, we may understand others by creating a similar response in our brain to that found in the other person's brain. Contagious yawning, under this account, is one extreme example of this more general and, normally, more subtle tendency. This chapter will attempt to unpack these claims and place them alongside traditional concepts in social and cognitive psychology, such as empathy and theory of mind. The chapter will also consider how these processes may be disrupted after brain injury and in people with autism.

The overarching question of the chapter is how do we understand the mental states of others? **Mental states** consist of knowledge, beliefs, feelings, intentions, and desires. The process of making this inference has more generally been referred to as **mentalizing**. The term is generally used in a theory-neutral way, insofar as it is used by researchers from a wide spectrum of views. It could be contrasted with the

KEY TERMS

Mental states
Knowledge, beliefs, feelings, intentions and desires.

Mentalizing
The process of inferring or attributing mental states to others.

It just takes one yawn to start other yawns off. How does this kind of simple contagion mechanism relate to empathy and theory of mind?

term ‘theory of mind’, which has essentially the same meaning but has tended to be adopted by those advocating a particular position, namely the notion that there is a special mechanism for inferring mental states. According to some researchers, this theory-of-mind mechanism cannot be reduced to general cognitive functions such as language and reasoning, or those involved in imitating. These arguments lie at the heart of the social neuroscience enterprise in that they raise important and divisive issues about the nature of the mental and neural processes that support social behavior and the extent to which they are related to other aspects of cognition.

WHAT IS SIMULATION THEORY?

Simulation theory is not strictly a single theory but a collection of theories proposed by various individuals (e.g. Gallese, 2001; Goldman, 2006; Hurley, Clark, & Kiverstein, 2008; Preston & de Waal, 2002). However, common to them all is the basic assumption that we understand other people’s behavior by recreating the mental processes on ourselves that, if carried out, would reproduce their behavior – that is, we use our own recreated (or simulated) mental states to understand, and empathically share, the mental state of others. Within this framework there are various ways in which this could occur. Gallagher (2007) broadly distinguishes between two: one could create an explicit, narrative-like simulation of another person’s situation and behavior in order to understand it; or when we see someone else’s behavior (e.g. their action, emotional expression) we may automatically, and perhaps unconsciously, activate the corresponding circuits for producing this behavior in our own brain. These latter versions of simulation theory tend to be intimately linked to the idea of mirror systems in which perception is tightly coupled with action.

EMPATHY AND SIMULATION THEORY

The word **empathy** is relatively modern, being little more than 100 years old. It was coined by Titchener (1909) from the German word *einfühlung* (Lipps, 1903) and originally referred to putting oneself in someone else’s situation. This would also go under the contemporary name of **perspective taking**. This section will first consider the various different ways in which the term empathy is used today, which reveals potentially important differences in the way that it may be accounted for.

KEY TERMS

Empathy
In the broadest sense, an emotional reaction to (or understanding of) another person’s feelings.

Perspective taking
Putting oneself in someone else’s situation.

Empathy as a multi-faceted concept

If one starts with the working definition of empathy introduced above (‘putting oneself in someone else’s situation’) it is clear that there are subtle, but potentially crucial, different ways in which this could be understood. Some of these are listed below and are an abridged version from Batson (2009):

1. Knowing another person’s internal state, including his or her thoughts and feelings.
2. Adopting the posture or matching the neural response of an observed other.

3. Having an emotional reaction to someone else's situation, although it need not be the same reaction.
4. Imagining how I would feel/react in that situation (i.e. given *my* personal history, traits, knowledge, beliefs).
5. Imagining how the other person would feel/react in that situation (i.e. given *their* personal history, traits, knowledge, beliefs).

The first three scenarios differ with respect to whether the knowledge/feeling is the same in self and other. Knowing about another person's internal state need not necessarily imply that the observer shares that state. This important consideration lies at the heart of some tests of theory of mind, specifically **false belief** tasks, but they are relevant to some conceptions of empathy too. The second sense in which empathy is used ('adopting the posture or matching the neural response of an observed other') is the one most closely linked with mirror systems, imitation, and contagion (emotional contagion, yawning contagion, etc.). For example, one might feel **personal distress** in response to someone else's suffering. The third sense in which the term empathy may be used differs from the second in that the person's response is not matched. For instance, one might feel a sense of **pity** to another's situation or **sympathy** towards someone who is suffering. These reactions are directed outwards (other-oriented) rather than being self-oriented (as in personal distress), and the response of the perceiver does not match that of the other person. The fourth and fifth notions of empathy relate more directly to the idea of perspective taking, but they differ in the degree to which they are self-oriented versus other-oriented. The fourth scenario ('imagining how I would feel/react in that situation') could be construed as a shallow attempt to empathize, in which the level of success is dependent on self–other similarity rather than a true understanding of the other.

Given these somewhat different conceptions of empathy, it is not surprising that there is no single agreed-upon measure of empathy. Theory-of-mind tests, discussed in detail below, normally involve assessments based on linguistic reasoning of the sort 'If X believes Y then how will he/she behave in situation Z?' Others use neural or bodily responses to seeing others in pain, for example, as a measure of empathy (e.g. Bufalari, Aprile, Avenanti, Di Russo, & Aglioti, 2007; Jackson, Meltzoff, & Decety, 2005).

KEY TERMS

False belief
A belief that does not correspond to current reality.

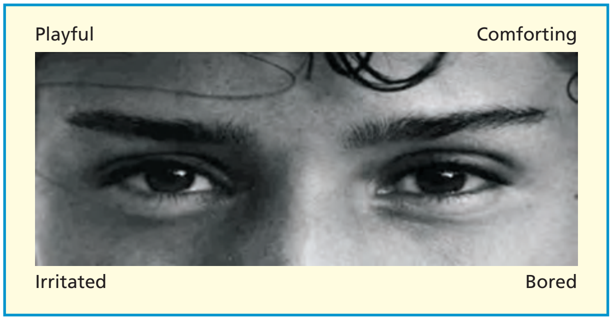
Personal distress
A feeling of distress in response to another person's distress or plight.

Pity
A concern about someone else's situation.

Sympathy
A feeling of compassion or concern for another person.

Do images of starvation evoke in you a sense of personal distress (self-focused) or a sense of pity or sympathy (other-focused)? Different individuals may have different reactions, although both can be broadly construed as empathic.

Of course, this presupposes a certain idea of what empathy is (i.e. that it can be measured solely in physiological ways). There are various questionnaire measures of empathy, such as the Interpersonal Reactivity Index (IRI;Davis, 1980) and the Empathy Quotient (EQ;Baron-Cohen & Wheelwright, 2004), which touch upon some of the distinctions discussed above. For example, the IRI contains separate subscales such as personal distress (items such as 'I tend to lose control during emergencies'), perspective taking (items such as 'Before criticizing somebody, I try to imagine how I would feel if I were in their place'), and empathic concern (items such as 'I often have tender, concerned feelings for people less fortunate than me'). One current trend is to incorporate questionnaire measures in functional imaging experiments. For example, watching someone drinking a pleasant or disgusting drink may activate the gustatory (taste) regions of the perceiver (Jabbi, Swart, & Keysers, 2007). Moreover, the extent to which this occurs may be greater in those people who report higher empathy on questionnaire measures (Jabbi et al., 2007). Findings such as these are often used to argue that the different concepts of empathy are related or, at least, share a common core (perhaps based upon simulation). Finally, one could potentially measure the ability to *accurately* empathize (i.e. to accurately state what another person is thinking or feeling) rather than the extent to which the person may report the motivation to empathize (i.e. most questionnaire measures) or to simulate that state themselves (which need not be linked to the ability to consciously report that state). As an example of such a test, the 'reading the mind in the eyes' test requires participants to match expressions in the eye region of faces to labels denoting mental states such as bored, sorry, or interested (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). Another test requires two participants to work together in a scenario that is video recorded. Each participant can then watch it back and report their own internal states as well as attempting to infer that of the other participant, thus enabling the experimenter to cross-reference the responses together in order to infer empathic accuracy (e.g. Ickes, 1993; Ickes, Gesn, & Graham, 2000). A recent functional imaging study based on this method found that empathic accuracy was related to a network of regions including the medial prefrontal cortex, implicated in mentalizing/theory of mind (although not the temporo-parietal junction), and the premotor cortex, which has been associated with mirror systems (Zaki, Weber, Bolger, & Ochsner, 2009).



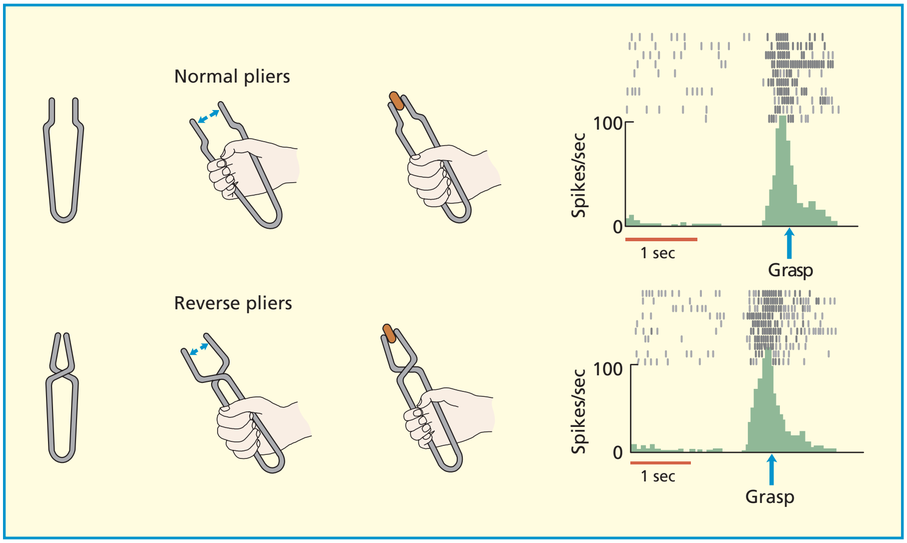
The extent to which people can accurately detect the mental states of others (also called empathic accuracy) may differ from the extent to which they try to empathize or take perspectives. One test along these lines is the 'reading the mind in the eyes' test. From Baron-Cohen et al. (2001). Copyright © 2001 Association for Child Psychology and Psychiatry. Reproduced with permission from Wiley-Blackwell.

From imitation to empathy?

A link between imitation and empathy receives some support from social psychology. These studies generally use unconscious imitation in which the participant engages in a task with another person (a confederate) and the extent to which the participant imitates the confederate is measured. The participant is unaware of the true nature of the study (i.e. that his/her imitative behavior is being assessed). Participants who imitate more (based on blind scoring of their actions) whilst performing a cooperative task with a confederate tend to rate themselves as higher in trait empathy (Chartrand & Bargh, 1999). When the confederate deliberately imitates the participant in a cooperative task, then he/she is liked more by the participant than in a control condition in which imitation is avoided (Chartrand & Bargh, 1999). Van Baaren, Holland, Kawakami and van Knippenberg (2004) showed that being imitated increases the chances of helping behavior when a confederate drops something. However, the effects are quite general. The person who has been imitated is not just more likely to help the imitator but they are more likely to help others too. It also increases the amount of money that the participant opts to donate to charity at the end of the experiment.

Iacoboni (2009) has argued that the mirror system for action may be co-opted by other regions of the brain to support empathy. Mirror neurons respond both when an animal performs an action and when it observes another performing the same (or similar) action – they act as a neural 'bridge' between self and other. They respond not just to the motor properties of an action but to the goal of the action. For example, it has been shown that neurons that respond to grasping respond in different ways to the sight of the grasp according to whether a container is present or absent (Fogassi et al., 2005). In this study the presence of the container was reliably associated with one particular goal, placing a piece of food inside it, whereas the absence was associated with another goal, eating it. In this example, the action is the same (grasp) but the subsequent goal is not and the mirror neurons (in the parietal lobe) respond according to the implied goal. Umilta et al. (2008) have shown that neurons that respond to grasping will also respond when pliers are used to grasp, even when a different action is required. In this example, the action is different but the goal is the same and the neural response is determined by the goal. Studies such as these have been used to argue that mirror neurons enable understanding of at least one mental state: intentions.

Carr, Iacoboni, Dubeau, Mazziotta, and Lenzi (2003) examined more directly a possible link between empathy and imitation using fMRI in humans. They showed



Mirror neurons respond to the same goal rather than the same action. Mirror neurons in monkeys responded similarly after training with both normal pliers and reverse pliers (maximum responding at point of grasping the food), even though both required different actions. From Umilta et al. (2008). Copyright © 2008 Proceedings of the National Academy of Science, USA. Reproduced with permission.

participants emotional facial expressions under two conditions: observation versus deliberate imitation. (Note that this is different from the social psychology studies above, in which imitation was spontaneous rather than instructed.) They found increased activation for the imitation condition relative to observation in classical mirror system areas such as the premotor cortex. In addition, they found increased activation in areas involved in emotion, such as the amygdala and insula. Their claim was that imitation activates shared motor representations between self and other but, crucially, there is a second step in which this information is relayed to limbic areas via the insula. This action-to-emotion route was hypothesized to underpin empathy. Other studies have reported a positive correlation between questionnaire-based empathy scores and activation in the premotor region when observing actions (Kaplan & Iacoboni, 2006) or listening to actions (Gazzola, Aziz-Zadeh, & Keysers, 2006).

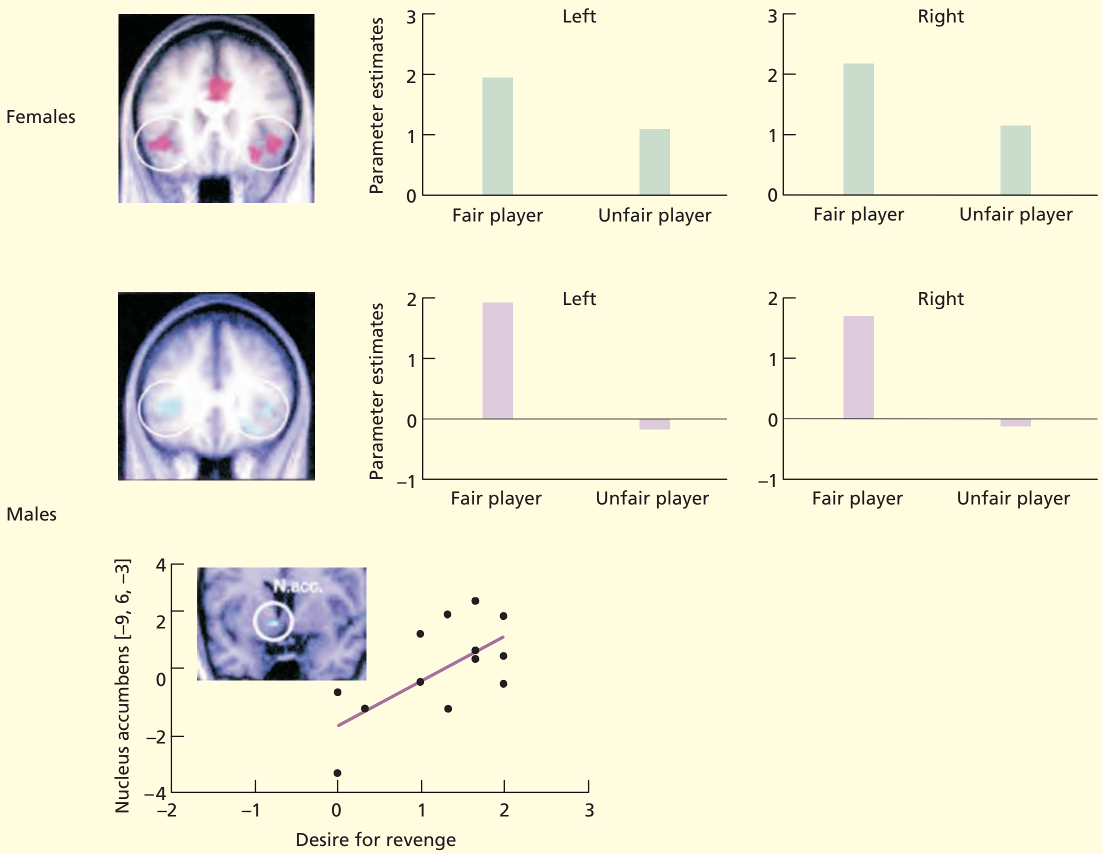
The model proposed by Carr et al. (2003) and Iacoboni (2009) is simple, but it is also perhaps simplistic. The assumption that limbic = emotion is an over-simplification (Le Doux, 1996), as is the claim that emotion imitation = empathy. As argued above, empathy is a broader concept than this. Recall also from Chapter 3 that the link between mirror neurons themselves and imitation is by no means uncontroversial. For example, monkeys (who posses mirror neurons) do not imitate tool use without extensive training.

It is possible to imagine alternative scenarios to the imitation-to-empathy model within a general simulation theory framework. For example, de Vignemont and Singer (2006) suggest that it may be possible to have simulation of emotions (and empathy for emotions) without having action/motor representations as a linking step. Singer, Seymour et al. (2004) investigated empathy for pain in humans using fMRI. The brain was scanned when anticipating and watching a loved-one suffer a mild electric shock. There was an overlap between regions activated by expectancy of another person's pain and experiencing pain oneself, including the anterior cingulate cortex and the insula. This provides evidence for a mirror system for pain – a system that responds to pain in self and other. However, there was little evidence that this system depends on the 'classic' mirror system for actions/goals that may support imitation.

Empathy beyond simulation

Some theories of empathy propose a variety of different mechanisms of which simulation is only one. In such models, simulation may either be a junior or senior partner.

As noted above, watching someone in pain activates certain parts of our own pain circuitry. This offers clear support for simulation theories. However, our beliefs about the person in pain can modulate or over-ride this mechanism. Singer et al. (2006) had participants in an fMRI scanner play a game with someone who plays fairly (a 'Goodie') and someone else who plays unfairly (a 'Baddie'). Mild electric shocks were then delivered to the Goodie and Baddie (who, of course, were only virtual characters but the participant did not know this). Participants empathically activated their own pain regions when watching the Goodie receive the electric shock. However, this response was attenuated when they saw the Baddie receiving the shock. In fact, male participants often activated their pleasure and reward circuits (such as the nucleus accumbens) when watching the Baddie receive the shock, which is the exact opposite of simulation theory. This brain activity correlated with their reported desire for revenge, which suggests that although simulation may tend to operate automatically it is not protected from our higher order beliefs.



Females (pink) and males (blue) show reduced activity in brain regions that respond to pain when watching an unfair player receive a shock (shown here for the insula). In males, activity in the nucleus accumbens, measured whilst the unfair player received a shock, correlates with their self-reported desire for revenge. From Singer et al. (2006). Copyright © 2006 Nature Publishing Group. Reproduced with permission.

The findings of this study have implications for conditions associated with a lack of empathy, such as autism and psychopathy. It suggests that there are multiple reasons why empathy might fail – because of a failure to simulate the emotions of others or because of personally or socially constructed beliefs about who is 'good' and who is 'bad'. The eminent social psychologist Bandura (2002) argues that simulation has a relatively minor role to play in empathy, arguing that if it did it would lead to emotional exhaustion, which would debilitate everyday functioning. Moreover, Bandura (2002) argues that acts of inhumanity, such as genocide, depend on our ability to self-regulate and dissociate self from other. Although genocide is an extreme example, displaying lack of empathy towards socially marginalized groups (e.g. illegal immigrants, welfare cheats) could be regarded as a typical facet of human behavior.

Other studies support this view. Although doctors may be expected to show empathy for their patients, it would be unhelpful for them to experience personal distress when performing painful procedures. Indeed acupuncturists show less

It may be important for doctors performing painful procedures to switch off their empathic tendencies. What kind of mechanisms in the brain might support this?

activity, measured by fMRI, in the pain network (including the anterior insula and anterior cingulate) when watching needles inserted into someone, relative to controls (Cheng et al., 2007). Lamm, Batson, and Decety (2007) found that activity in these pain-related regions, induced by watching painful facial expressions induced by medical treatment, was modulated by the observer's beliefs about whether the treatment was successful or not (more activity in pain-processing regions when less successful). It was also related to whether the participants were instructed to imagine the feelings of the patient or to imagine themselves to be in that situation (more activity in pain-processing regions when imagining self). This suggests that the tendency to simulate is moderated by cognitive control (e.g. based on our beliefs) and also our efforts to take different perspectives.

Studies of imitation also show that the extent to which two people imitate each other depends on the characteristics of the imitator and the person being imitated, as well as characteristics of the social situation (van Baaren, Janssen, Chartrand, & Dijksterhuis, 2009). This suggests that imitation-based simulation is flexible and context sensitive, taking into account information beyond perception–action links. For example, imitation is less likely when the confederate has a social stigma such as a facial scar or is heavily obese (Johnston, 2002). Similarly, non-deliberate imitation of facial expressions is greater for one's ethnic ingroup relative to an outgroup (Bourgeois & Hess, 2008).

Some models of empathy propose a divide between so-called cognitive empathy and affective empathy (e.g. Baron-Cohen & Wheelwright, 2004; Shamay-Tsoory, Aharon-Peretz, & Perry, 2009). For example, in the experiment of Singer et al. (2006) the tendency to simulate another's pain would be part of the affective empathy system, and the representation of the other's intentions (to deceive or cooperate) would be part of the cognitive empathy system (which is often linked to a theory of mind in general). The ability to regulate (e.g. inhibit) the affective responses evoked by seeing another in pain would also be linked to this system. The terms 'cognitive' and 'affective' require some clarification. Many researchers would not regard emotions as existing outside of cognition (Lazarus, 1984; Phelps, 2006). A better terminology might be affective and non-affective empathy, as this stresses the different informational content. Patients with acquired brain damage to the orbital and ventromedial prefrontal cortex have difficulties in recognizing emotions in others (Hornak et al., 1996) as well as reporting feeling less emotions in themselves (Hornak et al., 2003). These patients may fail tests of theory of mind based on affective information but not on non-affective information (Shamay-Tsoory, Tibi-Elhanany, & Aharon-Peretz, 2006). This provides some support for the affective/non-affective ('cognitive') distinction. However, strictly speaking it does not prove that there is a separate affective theory-of-mind 'module', only that this kind of affective reasoning task depends on the integrity of regions that give rise to our own emotional feelings.

Most simulation theories do not fit squarely in either of the putative 'cognitive' or affective divisions. For example, emotion contagion would be an example of simulation based on affective information, whereas studies on action and mirror neurons suggest that it is possible to simulate goals and intentions, which are ‘cognitive’ (i.e. non-affective) mental states. Mirror neurons themselves are non-affective insofar as their response does not differ between actions that result in a reward (e.g. grasping food) and those that do not (e.g. grasping an object) (Rizzolatti & Craighero, 2004).

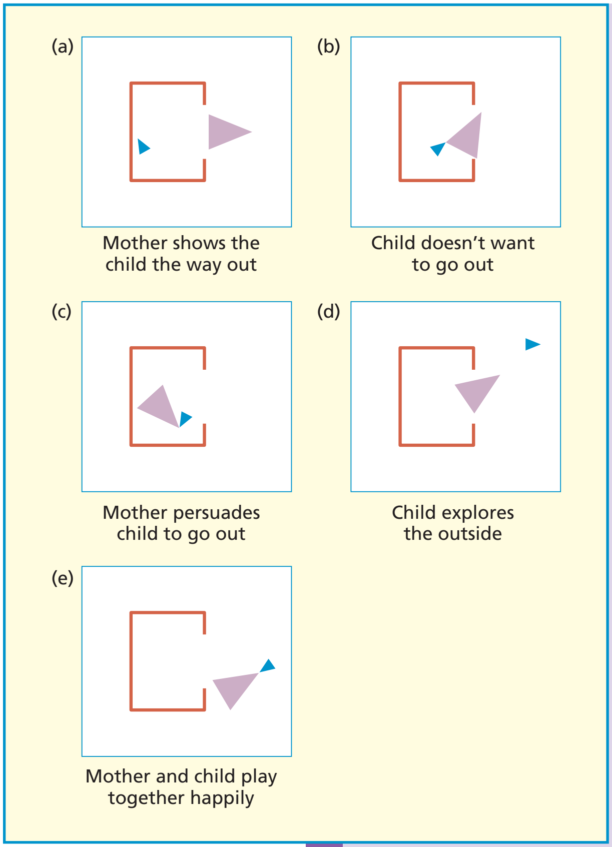
The model of empathy proposed by Decety and Jackson (2004, 2006) argues for a distinction between mechanisms based on simulation and other types of mechanism, but does not draw a sharp line between affective and non-affective processes. It brings together many of the strands discussed already. Decety and Jackson (2004) argue that there are three components of empathy:

1. *Shared representations between self and other, based on perception–action coupling.* This would include mechanisms for action understanding and imitation, emotional contagion, and pain processing. However, Decety and Jackson (2004) suggest that these are widely distributed throughout the brain rather than all loading on some core regions (such as premotor cortex).
2. *An awareness of self–other as similar but separate*. This is related to mechanisms of self-awareness (see Chapter 9) that enable us to attribute our own thoughts and actions as self-generated. Decety and Jackson (2004) suggest that one important brain region for this process is the right temporo-parietal junction (rTPJ). For instance, this region responds more when watching a moving dot controlled by someone else’s action relative to self-generated action (Farrer & Frith, 2002) and responds more when participants are asked to imagine someone else’s feelings and beliefs compared to their own (Ruby & Decety, 2004).
3. *A capacity for mental flexibility to enable shifts in perspective and self-regulation*. Decety and Jackson (2004) suggest that this is a candidate for a uniquely human component of empathy. It involves deliberate perspective taking of another’s situation, which may also involve inhibiting one’s own beliefs and self-referential knowledge. People with high self-reported personal distress may tend to over-rely on emotional contagion rather than cognitive control. Eisenberg et al. (1994) have shown that individual differences in personal distress are related to ability to control and shift attention, and Spinella (2005) reports negative correlation between behavioral measures of executive function and personal distress. Decety and Jackson (2004) suggest that regions in the prefrontal cortex responsible for the control of emotions (ventromedial and orbital regions) and the control of thought and action (lateral regions) are important. A region in the medial prefrontal cortex (considered below and in Chapter 9) responds to self-referential perspective relative to other perspective.

As such, this model offers a good account of the multi-faceted nature of empathy both in terms of cognitive mechanisms, social influences, and neural substrates. It also offers one way of connecting the literature on empathy with the other main topic of this chapter: theory of mind.

Evaluation

Empathy should perhaps best be regarded as a multi-faceted concept, and is likely to be explained via several interacting mechanisms rather than a single one. One possible division is between affective and cognitive (or non-affective) empathy, in which the former is based on emotion simulation and the latter on mental state reasoning.



Mother shows the child the way out

Child doesn't want to go out

Mother persuades child to go out

Child explores the outside

Mother and child play together happily

Mental states (e.g. want), behaviors (e.g. play), and other human characteristics (e.g. mother, child) are readily attributed to animated geometric shapes. Watching these animations, during functional imaging, activates a network of regions implicated in theory of mind. The captions were not presented in the studies, but are shown here for clarification. From Castelli et al. (2000). Copyright © 2000 Elsevier. Reproduced with permission.

PROJECTING MENTAL STATES EVERYWHERE – THE ORIGINS OF ANTHROPOMORPHISM?

Anthropomorphism refers to the attribution of human characteristics to non-human animals, objects, or other concepts. This could reflect a natural tendency to attribute mental states externally, and not just to other humans who are 'like me'. Living objects are commonplace in our popular culture – think of Pixar's bouncing lamp. It has also been suggested that a belief in God is a result of the tendency to attribute mental states externally (Guthrie, 1993).

To some extent, the tendency to anthropomorphize may depend on whether something looks like us – an angry dog shows its teeth like an angry human. Movement as well as appearance is important. Heider and Simmel (1944) found that people readily ascribe mental states to animations of two interacting geometric objects, such as 'the blue triangle wanted to surprise the red one'. In a functional imaging study that compared these kinds of animations with aimless movements, it was found that these moving shapes activated a network of regions that are typically activated in theory-of-mind tasks (Castelli, Happe, Frith, & Frith, 2000). They argued that this supports the idea that intentions tend to be inferred from actions, even in situations in which participants know that the objects are not capable of having mental states.

Although anthropomorphism may be a universal tendency, some people may do it more and others may do it less. One study found that this tendency, measured in terms of mental state ratings for gadgets or terms used to describe pets, is greater in lonely people (Epley, Akalis, Waytz, & Cacioppo, 2008). This suggests that it may be a compensatory mechanism for social isolation. In contrast, people with autism use less mental state terms to describe the moving geometric shape stimuli and show less activity in regions linked to theory of mind when watching these animations (Castelli, Frith, Happe, & Frith, 2002).

KEY TERMS

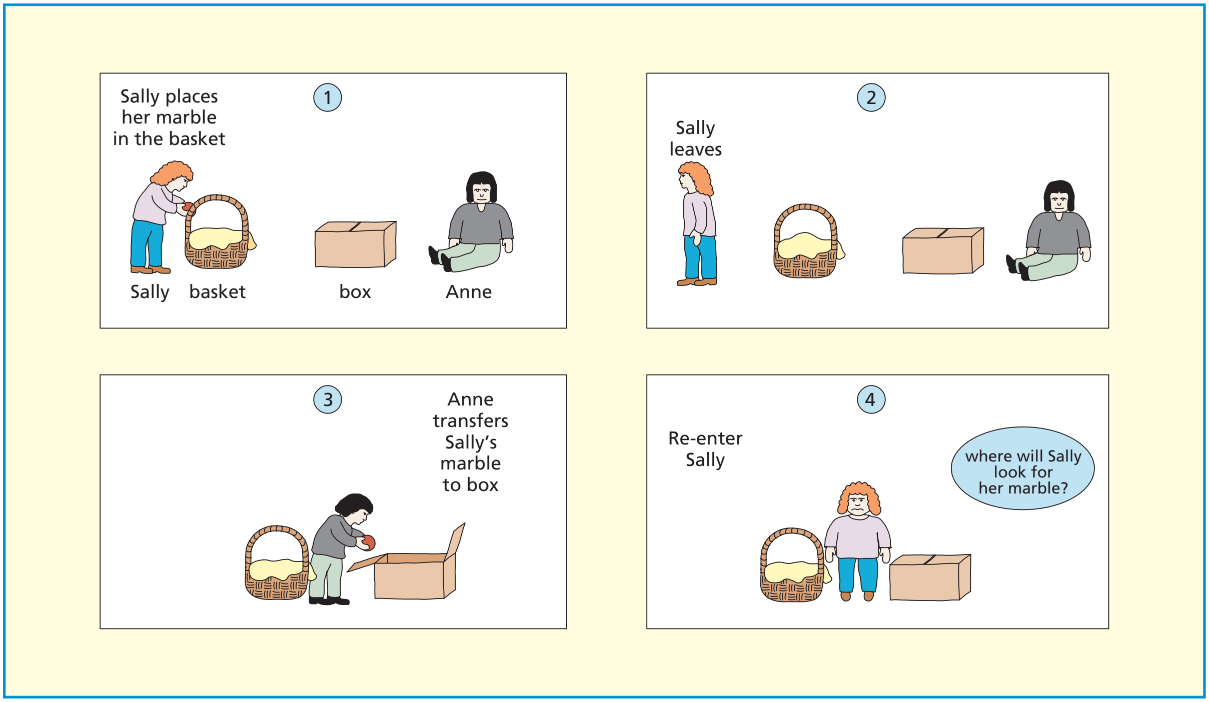
Anthropomorphism
The attribution of human characteristics to non-human animals, objects, or other concepts.

The model of Decety and Jackson proposes a set of different mechanisms that underpin empathy, but without evoking a dichotomy between cognitive/affective empathy. The idea of simulation is likely to remain an important component of models of empathy for the foreseeable future, but whether or not it is the main or core component of empathy in real-life social situations remains to be determined. Certainly, there is evidence that behaviors related to simulation, such as emotion contagion, are modulated by social biases, beliefs, and deliberate attempts at cognitive control (e.g. when deliberately adopting the other perspective).

THEORY OF MIND AND REASONING ABOUT MENTAL STATES

This section distinguishes itself from the previous one by considering in detail a certain kind of task: namely deliberate attempts to reason about mental states, and deliberate attempts to attribute mental states to others. To some extent these sorts of mechanisms are linked to those involving empathy, as discussed previously. However, the tasks used in the theory-of-mind literature are typically quite different from those considered previously in the section on empathy. The stimuli themselves are typically narratives or sequences of events, rather than observation of a particular state (e.g. pain). The tasks also typically require an overt response (e.g. what does Sally think or do?) whereas studies on empathy often do not (e.g. a typical measure could be degree of imitative behavior or subtle contraction of facial muscles). We may be able to tell from someone's face or voice that they are being thoughtful, but knowing what they are thinking may involve a different computation.

The term 'theory of mind' derived originally from research on primate cognition. Premack and Woodruff (1978) conducted a number of studies on a chimpanzee to see if it understood an experimenter's intentions. For example, the chimp might point to a picture of a key when an experimenter was locked in a cage, the inference being 'he wants to get out'. A number of criticisms were leveled at the study. For instance, it may reflect knowledge of object associations (e.g. between key and lock) rather than mental states. In a reply to the article, Dennett (1978) suggested that one way of testing for theory of mind would be to consider false beliefs, in which someone else may hold a mental state (e.g. a belief) that differs from one's own belief and from the current state of reality. In developmental psychology, the paradigmatic false belief test is the object transfer task, such as the Sally–Anne task (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1983). Sally puts a marble in a basket so that Anne can see. Sally then leaves the room, and Anne moves the marble to a box. When Sally enters the room, the participant is asked 'where will Sally look for the marble?' or 'where does Sally think the marble is?' A correct answer ('in the basket') is typically taken to indicate the presence of a theory of mind. An incorrect answer is potentially more problematic to interpret. It could imply a lack of theory of mind. However, one also has to rule out other factors such as language comprehension difficulties or a failure to inhibit a more dominant response (one's own belief). False beliefs are harder to accommodate within simulation theories because one's own belief is at odds with that attributed to the other person. This cannot be done by straightforward simulation involving shared self–other representations. It requires taking one's own mental states 'offline' and creating a hypothetical scenario different to current reality. So-called meta-representation and pretense is often regarded as a hallmark of theory-of-mind ability (Leslie, 1987).


1
Sally places her marble in the basket
Sally
basket
box
Anne
2
Sally leaves
3
Anne transfers Sally's marble to box
4
Re-enter Sally
where will Sally look for her marble?

The Sally–Anne task requires an understanding of false belief and an attribution of second-order intentionality. Adapted from Wimmer and Perner (1983).

Social psychologists use the term **attribution** to refer to the process of inferring the causes of people's behavior. The philosopher Dennett (1983) uses his own term of **intentional stance** to refer to our tendency to explain behavior in terms of mental states, which could otherwise be considered synonymous with mentalizing or theory of mind. However, Dennett (1983) has a particularly useful way of describing different levels of intentionality that might be used to account for behavior. For example, an observer might have to evoke zero-order intentionality to explain the behavior of an object, first-order intentionality to explain the behavior of some animals, and second-order intentionality to explain some human behavior.

- *Zero-order intentionality*. The assumption that an agent possesses no beliefs and desires. It responds to stimuli reflexively, such as producing a scream when frightened or running to evade a predator.
- ***First-order intentionality.*** The inference that an agent possesses beliefs and desires, but not beliefs about beliefs. It may produce a scream because it *believes* a predator is present or *wants* others to run away.
- ***Second-order intentionality***. The inference that an agent possesses beliefs about other people's beliefs. It may produce a scream because it wants

KEY TERMS

Attribution
In social psychology, the process of inferring the causes of people's behavior.

Intentional stance
The tendency to explain or predict the behavior of others using intentional states (e.g. wanting, liking).

First-order intentionality
An agent possesses beliefs and desires, but not beliefs about beliefs.

Second-order intentionality
An agent possesses beliefs about other people's beliefs.

others to believe that a predator is nearby. False belief tests operate at this level (e.g. 'I think that Sally thinks that the marble is in the box'). .

- *Third-order intentionality*. An agent possesses beliefs about other people's beliefs concerning beliefs about other people, such as 'I think that John thinks that Sally doesn't know where the marble is'.

In this taxonomy, first-order intentionality and above would constitute 'mentalizing', taking an 'intentional stance' or theory of mind (depending on one's preferred term). Second-order intentionality does not have a special status (from a theoretical point of view), but it has acquired a special status by virtue of the fact that most tests of theory of mind operate at this level because they are more stringent and cannot be solved by stating one's own beliefs.

Domain-general versus domain-specific accounts of theory of mind

Domain specificity is linked to the notion of modularity (Fodor, 1983). A cognitive mechanism, or brain region, can be said to be domain specific if it is specialized to process only one kind of information. Thus, a domain-specific theory-of-mind mechanism would be a process that is specialized for attributing mental states (Leslie, 1987). There are two dominant lines of evidence that have been brought to bear on this. Firstly, there is the question of whether there is a specific region of the brain that responds to reasoning about mental states but not other kinds of things. It is possible that such a mechanism could be distributed in several locations, or that only one of the regions in that network is truly domain specific. Secondly, one can look to see if there are specific impairments in mental state attribution but not in other domains. Most evidence related to this question has come from the developmental condition of autism (e.g. Baron-Cohen, 1995b) but other lines of research have addressed this question from the perspective of acquired brain damage (e.g. Samson, 2009).

Historically, explanations of theory of mind have fallen into two camps that are termed **theory-theory** and simulation theory. Theory-theory argues that we store, as explicit knowledge, a set of principles relating to mental states and how these states govern behavior (e.g. Gopnik & Wellman, 1992). In this sense, the 'theory' in theory of mind is like a mental rule-book for understanding others. This can be contrasted with simulation theory, which in one form would argue that perceptual-motor systems (rather than thinking and theorizing) are all that is needed for understanding others (e.g. Gallese & Goldman, 1998). When phrased in this way, it is reasonable to say that theory-theory makes more domain-specific assumptions whereas simulation theory can be considered a domain-general account. However, one needs to be cautious in dividing explanations into black and white dichotomies. For example, some versions of simulation theory argue that we do reason about mental states (rather than it being solely an outcome of perceptual-motor processes) but these versions are distinguished from theory-theory by making the claim that our own mental states form the foundation for understanding others (e.g. Mitchell, Banaji, & Macrae, 2005a). In a review of the neuroimaging literature on theory of mind, Apperly (2008) concludes that the strong division between simulation theory and theory-theory is no longer useful. Apperly (2008) argues instead that many of the concepts from social neuroscience research are likely to be more fruitful for understanding theory of mind, including: an understanding of how processing of self-related and other-related information is car-

KEY TERMS

Theory-theory
The idea that we store, as explicit knowledge, a set of principles relating to mental states and how these states govern behavior.

ried out; how both conscious beliefs and unconscious intuitions drive behavior; and so on. However, what such a 'third way' explanation will look like remains to be seen.

Stone and Gerrans (2006) argue against the notion of a domain-specific theory-of-mind mechanism and propose instead that the available data are more consistent with the notion of theory of mind arising out of the interaction of several different mechanisms (and not theory-theory either). This kind of explanation is in the spirit of the models of empathy discussed previously (Decety & Jackson, 2004). It is to be noted that Stone and Gerrans (2006) do not reject the idea of domain specificity per se. They claim that there are domain-specific mechanisms for detecting eye gaze, for example, and claim that deficits here could contribute to problems in theory of mind. Whilst the idea of a domain-specific mechanism for theory of mind is controversial, the idea that theory of mind requires basic competency in a number of domain-general mechanisms such as executive functions is not controversial, and a basic competency in language may be required for many tasks.

Language ability in typically developing children predicts success on a false belief task independently of age (Dunn & Brophy, 2005), and deaf children whose parents are non-native signers are delayed in passing such a task (Peterson & Siegal, 1995). This suggests that language is important for the development of theory of mind. Language may serve several functions: both a social, communicative role and also the acquisition of semantic knowledge of mental state words such as 'want' and 'think'. For example, children have to learn that these words denote concepts that are privately held (Wellman & Lagattuta, 2000). However, once a normal theory of mind is established it may not be dependent solely on language. Evidence for this assertion comes from brain-damaged patients with acquired **aphasia**. Apperly, Samson, Carroll, Hussain, and Humphreys (2006) report a single case study of a man with left hemisphere stroke who was impaired in many aspects of language, including syntax comprehension, but showed no impairments on non-verbal tests of theory of mind, including second-order inferences (X thinks that Y thinks).

Having sketched out the battle lines, the next section will go on to consider the neural substrates for theory of mind as evidenced from functional imaging (of neurologically normal adults) and neuropsychology (of brain-damaged adults). The following section will then consider autism in detail. Developmental issues will be covered specifically in Chapter 11.

Neural substrates of theory of mind

Evidence for the neural basis of theory of mind has come from two main sources: functional imaging studies of normal participants and behavioral studies of patients with brain lesions. Numerous tasks have been used, including directly inferring mental states from stories (e.g. Fletcher et al., 1995), from cartoons (e.g. Gallagher et al., 2000), or when interacting with another person (e.g. McCabe, Houser, Ryan, Smith, & Trouard, 2001a). A review and meta-analysis of the functional imaging literature was provided by Frith and Frith (2003), who identified three key regions involved in mentalizing.

Temporal poles

This region is normally activated in tasks of language and semantic memory. Frith and Frith (2003) suggest that this region is involved with generating **schemas** that specify

KEY TERMS

Aphasia
Deficits in spoken language comprehension or production, typically acquired as a result of brain damage.

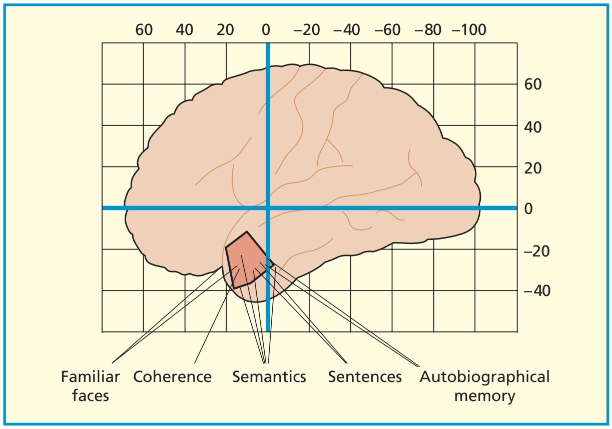
Schema
An organized cluster of different information (e.g. describing the subroutines of a complex action).

the current social or emotional context, as well as in semantics more generally. Zahn et al. (2007) report an fMRI study suggesting that this region responds to comparisons between social concepts (e.g. brave–honorable) more than matched non-social concepts (e.g. nutritious–useful). Also, not all the tests of mentalizing that activated this region involved linguistic stimuli. For example, one study used triangles that appeared to interact by, say, chasing or encouraging each other (Castelli et al., 2000).

Brain damage to the temporal poles is a feature of the degenerative disorder known as **semantic dementia** (Mummery et al., 2000). Patients with semantic dementia lose their conceptual knowledge of words and objects and show difficulties in language comprehension and production. However, there is little evidence from these patients that social concepts are selectively impaired. In general, although the temporal poles are important for theory of mind, there is no convincing support that it is domain specific for this kind of information.

KEY TERMS

Semantic dementia
A neurological condition associated with progressive deterioration in the meaning of objects and words.



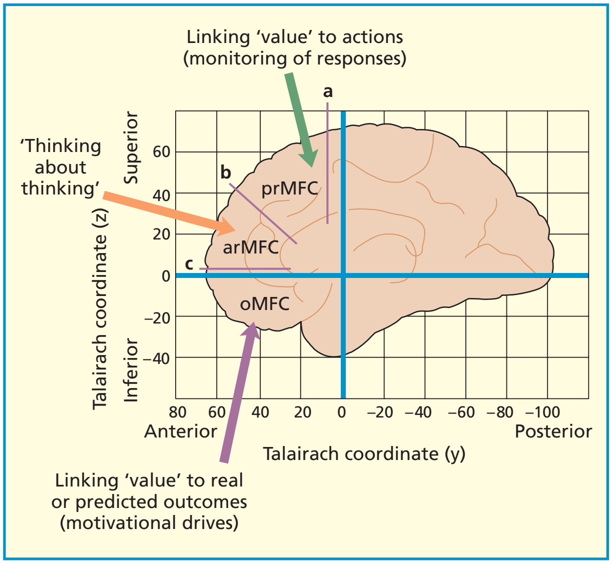
The temporal poles may support semantic knowledge, including of social concepts. Adapted from Frith and Frith (2003).

Medial prefrontal cortex (mPFC)

Frith and Frith (2003) reported that this region is activated in all functional imaging tasks of mentalizing to that date. Saxe (2006) argues that a sub-region of this area is involved in 'uniquely human' aspects of social cognition. This region lies in front of, but extends into, the ventral region of the anterior cingulate, labeled by Bush et al. (2000) as the affective division. Functional imaging studies reliably show that this region responds more to: thinking about people than thinking about other entities such as computers or dogs (e.g. Mitchell, Banaji, & Macrae, 2005b; Mitchell, Heatherton, & Macrae, 2002); thinking about the *minds* of people than thinking about their other attributes, such as their physical characteristics (Mitchell et al., 2005d); and thinking about the minds of certain people compared to others, such as similar people to ourselves (Mitchell et al., 2005b) and those who are most humanized relative to dehumanized (Harris & Fiske, 2006).

Some studies of patients with frontal lobe damage have suggested that the medial regions are necessary for theory of mind (e.g. Stuss, Gallup, & Alexander, 2001), but by no means all (e.g. Bird, Casteli, Malik, Frith, & Husain, 2004). This region also seems to be implicated in the pragmatics of language, such as irony ('Peter is well read. He has even heard of Shakespeare') and metaphor ('your room is a pigsty') (Bottini et al., 1994). Interestingly, people with autism have difficulties with this aspect of language (Happe, 1995). In such instances, the speaker's *intention* must be derived from the ambiguous surface properties of the words (e.g. the room is not literally a pigsty). Functional imaging suggests that this region is involved both in theory of mind and in establishing the pragmatic coherence between ideas/sentences, including those that do not involve mentalizing (Ferstl & von Cramon, 2002).

Can a generic function be ascribed to this region? If so, how does it relate to theory of mind? Amodio and Frith (2006) argue that the function of this region is in reflecting on feelings and intentions, which they label a 'meeting of minds'. One



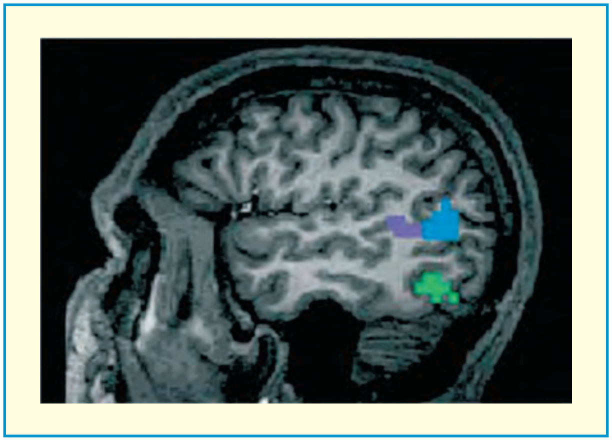
The medial frontal cortex (and adjacent regions of cingulate cortex) may contain three sub-regions with different functional specializations. Amodio and Frith (2006) regard the anterior rostral region as involved in 'thinking about thinking', or meta-cognition. This region is typically activated in tests of theory of mind. The orbital region is involved in linking value (positive or negative reinforcement) to outcomes, whereas the posterior rostral (or dorsal) region is involved in linking value to actions. These latter two regions are considered in Chapter 3. Adapted from Amodio and Frith (2006).

intriguing finding concerning this region is that it can be activated when a person believes they are playing a computer game against another person relative to when they think they are playing against a computer (Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004). Even though the situation is physically identical (the participant always played the computer), the act of cooperating with another person/mind engenders activity in this region. A more recent explanation of the function of this region is similar, but different, to that of Amodio and Frith (2006). Krueger, Barbey, and Grafman (2009) argue that the function of this region is to bind together different kinds of information (actions, agents, goals, objects, beliefs) to create what they term a 'social event'. They note that within this region some sub-regions respond more when participants make judgments about themselves and also about others who are considered to be similar to themselves (this is discussed in detail in Chapter 9). This suggests that this region is not attributing mental states per se, but is considering the self in relation to others (e.g. when playing a game against a human rather than a computer). It is also consistent with some versions of simulation theory in which participants understand others via deliberate perspective taking (e.g. Mitchell et al., 2005b). The notion of creating internal social events could also explain some of the findings of the role of this region in linking ideas in story comprehension (Ferstl & von Cramon, 2002).

Temporol-parietal junction (TPJ)

This region tends to be activated not only in tests of mentalizing but also in studies of the perception of biological motion, eye gaze, moving mouths, and living things in general. These skills are clearly important for detecting other 'agents' and processing their observable actions. Some simulation theories argue that mentalizing need not involve anything over and above action perception. It is also conceivable that this region goes beyond the processing of observable actions, and is also concerned with representing mental states and perhaps even the mental states of others over and above one's own mental states. Congenitally blind people activate essentially the same network of regions identified by Frith and Frith (2003) when they perform theory-of-mind tasks (Bedny, Pascual-Leone, & Saxe, 2009). This suggests that the computations of these regions are, at least partially, independent from visual perception of agents.

The TPJ region was previously highlighted in the discussion on empathy because it responds more when participants are asked to imagine how someone else would feel relative to how they would feel (e.g. Ruby & Decety, 2004). Patients with brain lesions in this region fail theory-of-mind tasks that cannot be accounted for by difficulties in body perception (Samson, Apperly, Chiavarino, & Humphreys, 2004). Saxe



According to Saxe (2006), the TPJ region may contain separate sub-regions for dealing with theory of mind (shown here in blue) and recognizing actions and expressions (shown here in purple). For comparison, the position of the extrastriate body area (in green) is shown, which is involved in body perception.

and Kanwisher (2003) found activity in this region, on the right, when comparing false belief tasks (requiring mentalizing) with false photograph tasks (not requiring mentalizing but entailing a conflict with reality). A false photograph may involve taking a picture of an apple on the tree, and then the apple falling down. In this scenario, there is a conflict between reality and a representation of reality. The result was also found when the false photograph involved people and actions, consistent with a role in mentalizing beyond any role in action/person perception. The region responds to false beliefs more than false maps or signs, which differ in an important way from a false photograph in that they are designed to represent *current* reality (Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006). Saxe and colleagues do not dismiss the fact that this region has a role to play in recognizing people and actions, but they claim that there may be different sub-regions within it, with one sub-region specialized for the attribution of mental states (Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009). Moreover, Saxe (2006) argues that it is uniquely human in doing so. It is important to note that this region is not specialized for false belief per se. It responds to true beliefs and other types of mental state (Saxe & Wexler, 2005). In other words, it responds to attributions of first-order intentionality as well as higher order intentionality (in Dennett's terms). Saxe and Powell (2006) have shown that this region responds to attribution of contentful mental states (such as thoughts and beliefs) rather than subjective states (such as hunger or tiredness). This suggests that it may have a role over and above 'thinking about others'. However, it is important to mention that one should be cautious in making strong claims about relative differences in BOLD signal. The differences can reflect different functional specialization (Saxe's claim) but they can also reflect the different difficulty of tasks, and the attention or strategy deployed to solve them. One could defend the claim of functional specialization by noting that other regions that respond to theory of mind do not show the same selective responses as the TPJ (Saxe & Wexler, 2005).

Evaluation

Functional imaging studies of the general population and, to a lesser extent, studies of people with acquired brain damage have helped to reveal the key regions involved in theory of mind and their somewhat different functions. There remains no consensus as to whether there is a domain-specific mechanism for theory of mind (i.e. a particular neural region that is dedicated to attributing mental states), but the strongest candidate region for domain specificity has shifted away from the medial prefrontal area to the TPJ region.

EXPLAINING AUTISM

> *He wandered about smiling, making stereotyped movements with his fingers, crossing them about in the air. He shook his head from side to side, whispering*

or humming the same three-note tune. He spun with great pleasure anything he could seize upon to spin ... When taken into a room, he completely disregarded the people and instantly went for objects, preferably those that could be spun ... He angrily shoved away the hand that was in his way or the foot that stepped on one of his blocks.

(This description of Donald, aged 5, was given by Leo Kanner (1943), who also coined the term autism. The disorder was independently noted by Hans Asperger (1944), whose name now denotes a variant of autism.)

Autism has been formally defined as 'the presence of markedly abnormal or impaired development in social interaction and communication and a markedly restricted repertoire of activities and interests' (American Psychiatric Association, 1994). It is a severe developmental condition that is evident before 3 years of age and lasts throughout life. There are a number of difficulties in diagnosing autism. First, it is defined according to behavior because no specific biological markers are known (for a review, see Hill & Frith, 2003). Second, the profile and severity may be modified during the course of development. It can be influenced by external factors (e.g. education, temperament) and may be accompanied by other disorders (e.g. attention deficit and hyperactivity disorder, psychiatric disorders). As such, autism is now viewed as a spectrum of conditions spanning all degrees of severity. It is currently believed to affect 1.2% of the childhood population, and is three times as common in males (Baird et al., 2006). **Asperger's syndrome** falls within this spectrum, and is often considered a special sub-group. The diagnosis of Asperger's syndrome requires that there is no significant delay in early language and cognitive development, although the term is also used to denote people with autism who fall within the normal range of intelligence. Learning disability, defined as an IQ lower than 70, is present in around half of all cases of autism (Baird et al., 2006).

Much of the behavioral data has been obtained from high-functioning individuals in an attempt to isolate a specific core of deficits. On a purely theoretical level, one reason why researchers have been interested in the study of autism is the belief that it might reveal something fundamental about social interactions more generally.

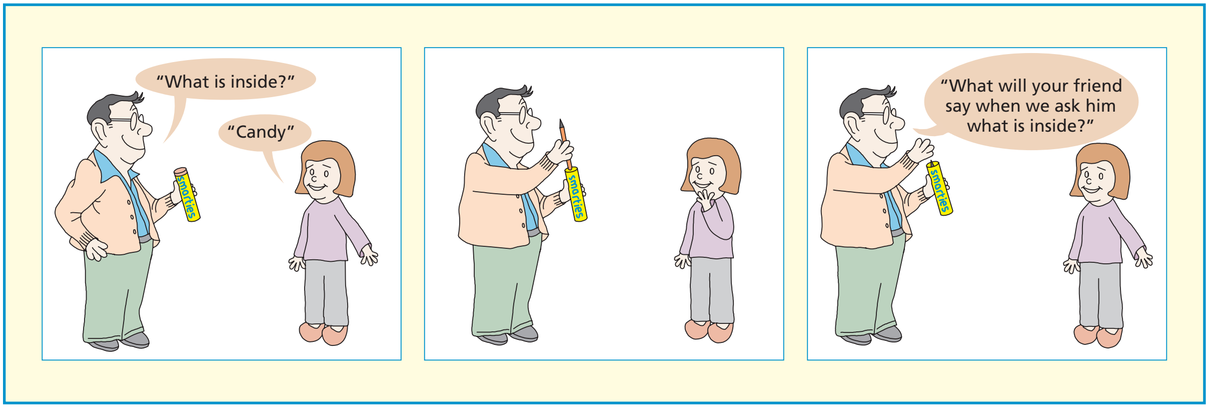
KEY TERMS

Autism
A developmental condition associated with the presence of markedly abnormal or impaired development in social interaction and communication and a markedly restricted repertoire of activities and interests.

Asperger's syndrome
A sub-type of autism associated with less profound non-social impairments.

Autism as mind blindness

One candidate deficit is the ability to represent mental states, or theory of mind (e.g. Baron-Cohen, 1995b; Fodor, 1992). The first empirical evidence in favor of this hypothesis came with the development of a test of false belief devised by Wimmer and Perner (1983) and tested on autistic children by Baron-Cohen et al. (1985) as the Sally–Anne task (described above). Autistic children tend to fail the task whereas normally developing children (from 4 years on) pass the test, as do control participants with learning disability matched in IQ to the autistic children. The erroneous reply is not due to a failure of memory, because the children can remember the initial location. It is as if they fail to understand that Sally has a belief that differs from physical reality – that is, a failure to represent mental states. This has also been called 'mind-blindness' (Baron-Cohen, 1995b). Autistic children are still impaired when the false belief was initially their own. For example, in one task, the child initially expects to find candy in a candy packet and is surprised to find a pencil, but



The child initially expects to find candy in a tube of Smarties and is surprised to find a pencil. When asked what other people will think is in the packet, autistic children reply 'pencil' whereas typically developing children reply 'candy'.

when asked what other people will think is in the packet the child replies 'pencil' (Perner, Frith, Leslie, & Leekam, 1989).

Passing false belief tasks requires the ability to form meta-representations (i.e. representations of representations: in this instance, beliefs about beliefs). It was originally suggested that a failure of meta-representation may account for impaired theory of mind in autism (Baron-Cohen et al., 1985). However, other studies suggest that autistic people can form meta-representations in order to reason about false photographs in which the information depicted on the photograph differs from current reality (Leekam & Perner, 1991). If their deficit really is related to mental state representations rather than physical representations, then this offers support for the domain-specific account. A number of other studies have pointed to selective difficulties in mentalizing compared to carefully controlled conditions. For example, people with autism can sequence behavioral pictures but not mentalistic pictures (Baron-Cohen, Leslie, & Frith, 1986); They are good at sabotage but not deception – they tend to think that everyone tells the truth (Sodian & Frith, 1992); and they tend to use desire and emotion words but not belief and idea words (Tager-Flusberg, 1992). In all instances, the performance of people with autism is compared to mental-age controls to establish that the effects are related to autism and not to general level of functioning.

Functional imaging studies of autistic people carrying out theory of mind (Happe et al., 1996) or related tasks (Castelli et al., 2002) have shown reduced activity in the network of regions commonly activated by controls.

Finally, it may be necessary to make a distinction between implicit mentalizing (intuitive, reflexive) and more explicit forms of mentalizing (based on reasoning). Whilst the latter tend to be measured by overt predictions of behavior, the former may be measured by non-declarative means (e.g. monitoring of eye movements). For example, some high-functioning people with autism pass standard theory-of-mind measures but may still lack an intuitive understanding of others and may still show abnormal performance on other measures (e.g. eye movements to a location consistent with a false belief; Senju, Southgate, White, & Frith, 2009). By contrast, children under the age of 4 years show some implicit understanding of false beliefs (based on the same measure) despite failing on explicit measures (Onishi & Baillargeon, 2005). This is considered in detail in Chapter 11.

Autism as executive dysfunction

The mentalizing or theory-of-mind account of autism has not been without its critics. These criticisms generally take two forms: that other explanations can account for the data without postulating a difficulty in mentalizing (e.g. Russell, 1997); or that a difficulty with mentalizing is necessary but insufficient to explain all of the available evidence (e.g. Frith, 1989). A number of studies have argued that the primary deficit in autism is one of executive functioning (Hughes, Russell, & Robbins, 1994; Ozonoff, Pennington, & Rogers, 1991; Russell, 1997). **Executive functions** refer to control processes that are needed to coordinate the operation of more specialized components of the brain, thus, enabling us to switch attention from one task to another, to give priority to certain kinds of information, or to develop novel solutions, which would include inhibiting familiar solutions (e.g. Goldberg, 2001). For example, the incorrect answer might be chosen on false belief tasks because of a failure to suppress the strongly activated 'physical reality' alternative. Some patients with brain damage in prefrontal regions do this when given false belief tasks (Samson, 2009). However, it is not clear that this explanation can account for all the studies relating to mentalizing (e.g. picture sequencing). Moreover, high-functioning autistic people often have normal executive functions (e.g. Baron-Cohen, Wheelwright, Stone, & Rutherford, 1999) and early brain lesions can selectively disrupt theory-of-mind abilities without impairing executive functions (e.g. Fine, Lumsden, & Blair, 2001).

KEY TERMS

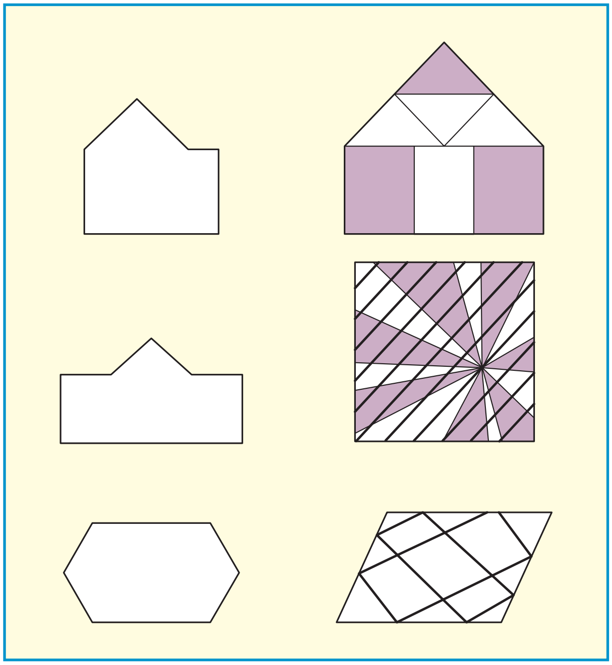
Executive functions
Control processes that are needed to coordinate the operation of more specialized components of the brain, thus enabling us to switch attention from one task to another, to give priority to certain kinds of information, or to develop novel solutions, which would include inhibiting familiar solutions.

Rather than difficulties in executive function explaining impairment on theory-of-mind tasks, Baron-Cohen (2009) speculates that the opposite could be true – namely, autistic people may develop, and stick to, their own rule system rather than the 'correct' one as determined by another person, the experimenter. An experiment is, in effect, a social contract. One study found that autistic people show the greatest impairment on open-ended tasks of executive function (in which participants may induce their own rules), rather than those that require the following of simple, stated rules (White, Burgess, & Hill, 2009). On some tests of executive function autistic people show differences in the medial prefrontal region, which is implicated in mentalizing (Gilbert, Bird, Brindley, Frith, & Burgess, 2008). This again suggests that difficulties on some aspects of executive functions could be related to their social difficulties.

Autism as weak central coherence

One difficulty with the theory-of-mind explanation is that it fails to account for cognitive strengths as well as weaknesses. One popular notion of autistic people is that they have unusual gifts or 'savant' skills, as in the film *Rain Man*. In reality, these skills are found only in around 10% of the autistic population (Hill & Frith, 2003). Nevertheless, some account of them is needed for a full explanation of autism. The unusual skills of some autistic people may be partly an outcome of their limited range of interests. Perhaps one reason why some individuals are good at memorizing dates is that they practice it almost all the time. However, there is also evidence for more basic differences in processing style. For example, people on the autistic spectrum are superior at detecting embedded figures (Shah & Frith, 1983) and searching for a target in an array of objects (for a review see Mitchell & Ropar, 2004). One explanation for this is in terms of 'weak central coherence' (Frith, 1989; Happe, 1999). This is a cognitive style, assumed to be present in autism, in which processing of parts (or local features) takes precedence over processing of wholes (or global features).

What would cause such a pattern? One study has suggested that weak central coherence is linked to differences in brain size and connectivity (White, O'Reilly, & Frith, 2009). However, it is also possible that differences in social cognition in autism cause differences in the style of perceptual processing, rather than vice versa. For example, cultures that regard themselves as socially inter-dependent (i.e. strongly connected with the people around them in terms of shared goals an identity) show more global processing than those who construe themselves more socially independent (Davidoff, Fonteneau, & Fagot, 2008; Lin & Han, 2009; Nisbett, Peng, Choi, & Norenzayan, 2001). People with autism could be regarded as lying at one extreme end of this normal scale.



People with autism may be faster at spotting embedded figures such as the ones shown here (the figures on the left are embedded within those on the right).

Autism as an extreme form of the male brain

Baron-Cohen (2002, 2009) argues that the characteristics of all individuals can be classified according to two dimensions: 'empathizing' and 'systemizing'. Empathizing allows one to predict a person's behavior and to care about how others feel. Systemizing requires an understanding of lawful, rule-based systems and requires an attention to detail. Males tend to have a brain type that is biased towards systemizing (S > E) and females tend to have a brain type that is biased towards empathizing (E > S). However, not all men and women have the 'male type' and 'female type', respectively. Autistic people appear to have an extreme male type (S >> E), characterized by a lack of empathizing (which would account for the mentalizing difficulties) and a high degree of systemizing (which would account for their preserved abilities and unusual interests). Questionnaire studies suggest that these distinctions hold true (Baron-Cohen, Richler, Bisarya, Gurunathan, & Wheelwright, 2003; Baron-Cohen & Wheelwright, 2004). However, it remains to be shown whether these distinctions are merely descriptive or indeed do reflect two real underlying mechanisms at the cognitive or neural level.

How does the extreme male brain hypothesis relate to other theories of autism? Baron-Cohen (2002, 2009) regards this explanation as an extension of the earlier mind blindness theory, which has the advantage of being able to incorporate additional data. Specifically, it accounts for some of the non-social differences found in autism and it offers an explanation for why autism is more common in men (i.e. because men are more likely to have S>E type brains). However, there are at least

SYSTEMIZING IN CLASSIC AUTISM AND/OR ASPERGER'S SYNDROME

Type of systemizing	Classic autism	Asperger's syndrome
sensory systemizing	tapping surfaces or letting sand run through one's fingers	insisting on the same foods each day
motoric systemizing	spinning round and round, or rocking back and forth	learning knitting patterns or a tennis technique
collectible systemizing	collecting leaves or football stickers	making lists and catalogues
numerical systemizing	obsessions with calendars or train timetables	solving maths problems
motion systemizing	watching washing machines spin round and round	analysing exactly when a specific event occurs in a repeating cycle
spatial systemizing	obsessions with routes	developing drawing techniques
environmental systemizing	insisting on toy bricks being lined up in an invariant order	insisting that nothing is moved from its usual position in the room
social systemizing	saying the first half of a phrase or sentence and waiting for the other person to complete it	insisting on playing the same game whenever a child comes to play
natural systemizing	asking over and over again what the weather will be today	learning the Latin names of every plant and their optimal growing conditions
mechanical systemizing	learning to operate the VCR	fixing bicycles or taking apart gadgets and reassembling them
vocal/auditory/verbal systemizing	echoing sounds	collecting words and word meanings
systemizing action sequences	watching the same video over and over again	analysing dance techniques

Source: Baron-Cohen, S., Ashwin, E., Ashwin, C., Tavassoli, T., & Chakrabarti, B. (2009). Talent in autism: Hyper-systemizing, hyper-attention to detail and sensory hypersensitivity. *Philosophical Transactions of the Royal Society of London, Series B*, *364*(1522), 1377–1383. Copyright © 2009 The Royal Society. Reproduced with permission.

two ways in which these different ideas (mind blindness vs extreme male brain) could be related: that an inability to engage with others (due to a theory-of-mind deficit) leads to systemizing as a kind of compensatory strategy; or that an unusual interest or ability in systemizing leads to a lack of interest and understanding of social behavior. A third possibility is that both are true – that whatever it is that causes high systemizing also causes low empathizing. Possible mechanisms include fetal testosterone levels (e.g. Auyeung et al., 2009) or sex-related genetic differences (e.g. Creswell & Skuse, 1999). Although the extreme male brain theory predicts an autistic advantage for understanding systems, it differs from the weak central coherence theory by not making predictions about a difference between local versus global information. Finally, some research has tried to suggest a link between the extreme male brain theory and the broken mirror theory (discussed below), noting that there are sex differences (within the non-autistic population) in white/gray matter density in regions associated with the mirror system, with females showing greater density (Cheng et al., 2009). An EEG signature linked to functioning of the mirror system, termed **mu suppression**, also shows a sex difference, with females showing greater suppression (Cheng et al., 2008).

KEY TERMS

Mu suppression
The tendency for fewer mu waves (in EEG) to be present during the execution of an action.

The broken mirror theory of autism

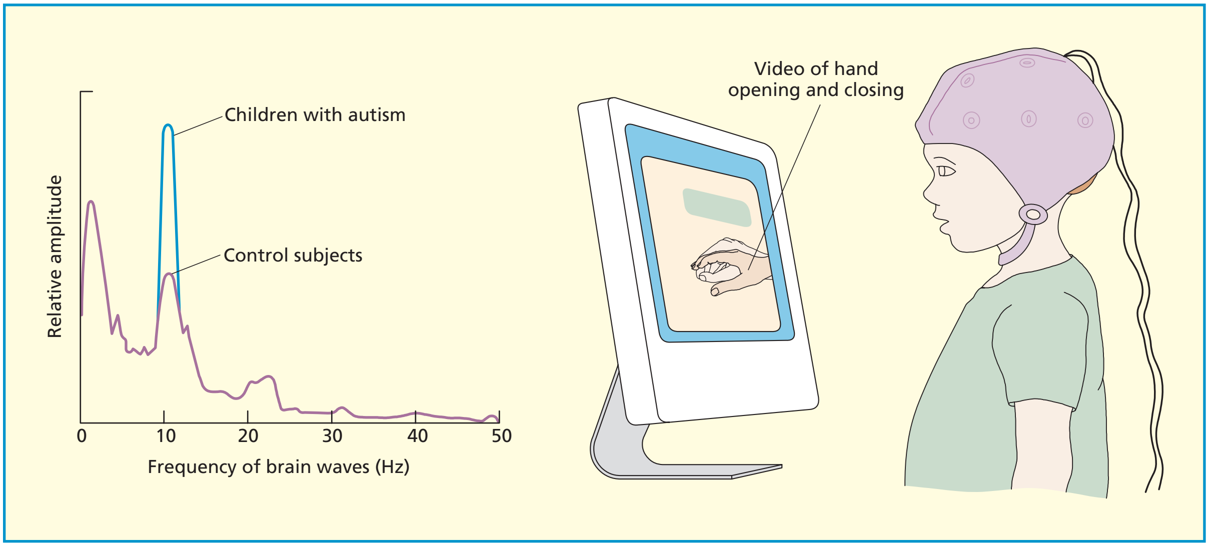
The **broken mirror theory** of autism argues that the social difficulties linked to autism are a consequence of mirror system dysfunction (Iacoboni & Dapretto, 2006; Oberman & Ramachandran, 2007; Ramachandran & Oberman, 2006; Rizzolatti & Fabbri-Destro, 2010). Hadjikhani, Joseph, Snyder, and Tager-Flusberg (2006) examined, using structural MRI, the anatomical differences between the brains of autistic individuals and matched controls. The autistic individuals had reduced gray matter in several regions linked to the mirror system, including the inferior frontal gyrus (Broca's region), the inferior parietal lobule, and the superior temporal sulcus. Although these were not the only regions where differences were found, the degree of thinning in these regions correlated with autistic symptom severity.

EEG, fMRI, and TMS data also suggest differences in mirror system functioning during certain tasks. Oberman et al (2005) used EEG to record mu waves over the motor cortex of high-functioning autistic children and controls. **Mu waves** occur at a

KEY TERMS

Broken mirror theory
An account of autism in which the social difficulties are considered as a consequence of mirror system dysfunction.

Mu waves
EEG oscillations at a particular frequency (8-13 Hz) that are greatest when participants are at rest.



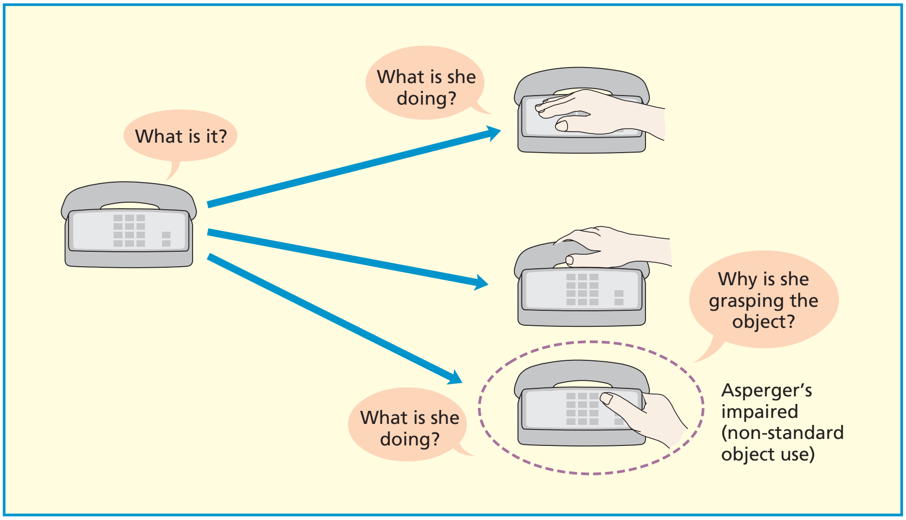
Mu waves are EEG oscillations in the 8–13 Hz range that are reduced both when performing an action and when watching someone else perform an action (relative to rest). As such, they may provide a neural signature for human mirror neurons. Autistic children show less mu suppression when watching others perform a hand action, which provides evidence in support of broken mirror theory. From Ramachandran, V. S., & Oberman, L. M. (2006). Broken mirrors: A theory of autism. *Scientific American*, 295(5), 62–69. Copyright © 2006 Scientific American. Reproduced with permission.

particular frequency (8-13 Hz) and are greatest when participants are doing nothing. However, when they perform an action there is a decrease in the number of mu waves, a phenomenon termed mu suppression. Importantly, in typical controls mu suppression also occurs when people *observe* actions and, as such, it has been regarded by some as a measure of mirror system activity (e.g. Pineda, 2005). Oberman et al. (2005) found that the autistic children failed to show as much mu suppression as controls during action observation (watching someone else make a pincer movement) but did so in the control condition of action execution (they themselves make a pincer movement).

Similar findings have been obtained with fMRI. Dapretto et al. (2006) conducted a study in which autistic children and matched controls either observed or imitated emotional expressions. The imitation condition produced less activity in the inferior frontal gyrus of the autistic children relative to controls, and this was correlated with symptom severity. Differences in regions linked to face recognition (fusiform gyrus) and emotion recognition (amygdala) did not differ between groups.

Finally, watching someone perform an action increases one's own motor excitability, measured as a motor-evoked potential (MEP) on the body when TMS is applied to the motor cortex. However, this effect is reduced in autistic people even though their motor cortex behaves normally in other contexts (Theoret et al., 2005).

The broken mirror theory makes some novel predictions about what people with autism might be impaired at, such as imitation and understanding the goals of others based on action observation. Boria et al. (2009) compared children with autism against typically developing controls in which actions were either


What is it?
What is she doing?
Why is she grasping the object?
What is she doing?
Asperger's impaired (non-standard object use)

People with autism perform worse than controls at inferring intentions for non-standard actions. In this example, they are more likely than controls to say that the person intends to make a call than to say that they are moving the phone. Figure based on Boria et al. (2009).

consistent with typical use of a phone (e.g. making a call) or not (e.g. picking it up to move it). They found that the autistic children were more likely to base their understanding of actions based on the object rather than action. In this example, they are more likely to answer ‘making a call’ when the object is being moved.

Although deficits in imitation are found in autism (Williams, Whiten, & Singh, 2004), these may be more apparent in spontaneous imitation than instructed imitation (e.g. Hamilton, Brindley, & Frith, 2007). This suggests that autistic people have a poor intuitive understanding of when and what to imitate (i.e. social rules) rather than in perceptual-motor interactions (at the heart of the broken mirror theory). The broken mirror theory has its critics (Dinstein, Thomas, Behrmann, & Heeger, 2008; Southgate & Hamilton, 2008). In general, the criticism takes too forms. Firstly, it does not account for all the unusual behavior found in autism (e.g. embedded figures; interest in systems). Defendants of the theory argue that it is not trying to explain all the features of autism (i.e. it is not a theory of autism but a theory of certain characteristics of autism). The second general criticism surrounds the extent to which empathy and imitation are linked to mirror systems. Earlier in the chapter, many examples were given of how both imitation and empathy are modulated by social rules, deliberate attempts at perspective taking, and so on. A core deficit elsewhere (e.g. in representing mental states) could nevertheless affect the functioning of the mirror system, and perhaps even lead to structural changes within that system. Heyes (2010) argues that the properties of mirror neurons may be learned as a result of social interactions. Thus, impoverished social interactions may cause mirror system dysfunction, as well as vice versa.

Evaluation

For many years the dominant explanation of autism has been that it fails to represent the mental states of others. This has been termed mind blindness and has tended to have been regarded as a failure to develop a theory of mind (although not necessarily with commitment to the idea that this exists as a domain-specific module). Other theories, such as weak central coherence theory and extreme male brain theory, maintain this basic idea but adopt a wider perspective in order to explain other features of autism. The most significant challenge to this idea previously came from the notion of executive dysfunction in autism, but now comes in the form of broken mirror theory. There is good evidence of mirror neuron dysfunction in autism, but it is less clear whether this dysfunction is a core feature of autism or a by-product of other deficits – given that mirror systems in general are modulated by beliefs, social knowledge, and cognitive control.

SUMMARY AND KEY POINTS OF THE CHAPTER

- Simulation theory argues that we understand the mental states (thoughts, feelings, beliefs, etc.) of others by activating our own mechanisms for producing that behavior. To some extent, we literally share the experiences of the people around us. As such, simulation theory is an appealing way of explaining empathy.
- Empathy is a broad concept that may include simulation, but it is unlikely to be limited to it. It also involves perspective taking (either automatically or deliberately) and cognitive control, which may inhibit the tendency to simulate.
- Both empathy and theory of mind (or mentalizing) involve understanding the mental states of other, but the latter is typically assessed via conscious attempts to reason about mental states, such as in false belief tasks.
- Functional imaging of the normal population reveals a network of regions that are consistently activated by tests of theory of mind, and the two regions that have provoked the most research interest are the temporol-parietal junction region and a medial prefrontal cortex region. However, it remains controversial whether either region can be classed as domain specific for attributing mental states.
- People with autism often fail theory-of-mind tasks, leading to the theory that they have a specific impairment in representing mental states. Their difficulty is not well explained by difficulties in executive function alone or difficulties in meta-representation per se.
- There is good evidence of mirror neuron dysfunction in autism, but it is less clear whether this dysfunction is a core feature of autism or a by-product of other deficits (given that mirror systems in general are modulated by beliefs, social knowledge, and cognitive control).

EXAMPLE ESSAY QUESTIONS

- What is the evidence for and against simulation theories of empathy?
- How is empathy related to theory of mind, and in what ways are they different?
- Is there a theory of mind module in the human brain?
- How can the social behavior of people with autism be explained?

RECOMMENDED FURTHER READING

- Decety, J., & Ickes, W. (2009). *The social neuroscience of empathy*. Cambridge, MA: MIT Press. An excellent collection of papers on empathy.
- Hill, E. L., & Frith, U. (2004). *Autism: Mind and brain*. Oxford: Oxford University Press.
- Saxe, R., & Baron-Cohen, S. (2006). *Theory of mind*. New York: Psychology Press. A very good collection of papers on all aspects of theory of mind.