11

The role of the frontal lobes and the amygdala in theory of mind

VALERIE E. STONE

INTRODUCTION

One male macaque approaches another who is next to a food source. The one approaching moves his head forward, raises his brows, and opens his mouth in a threat display. The other monkey withdraws from the food source, signalling his submission by smacking his lips.

A graduate student is at a Society for Neuroscience conference, talking to a relatively well-known researcher at a poster, when the president of the society comes up. The president greets the researcher. The graduate student excuses himself, ducks his head and discreetly withdraws.

In both cases, a subordinate withdraws when a dominant animal approaches a valuable resource. The subordinate macaque, however, may simply respond to the dominant macaque's approach, without explicitly representing the dominant animal's mental state. He recognizes which other individual he is interacting with, remembers that this male is dominant to him, recognizes the facial gesture, and has a physiological and emotional response which leads him to withdraw and signal submission. The graduate student's response doubtless also includes this kind of emotional and physiological component, but he may also represent the researcher's and the president's mental states 'They want to talk to each other; she would probably rather talk to the president,' in order to compute the appropriate response.

This example illustrates some parallels between social cognition in humans and in other primates. Our social behaviour is similar in many ways to other primates'. We negotiate social hierarchies, keep track of kin members, compete over limited resources, and encode past histories of interaction with specific individuals. Humans, however, appear to have more complex abilities to represent others' mental states and to use these mental states as the basis of social computations. The term 'theory of mind' is used to refer to humans' ability to make inferences about others' mental states (e.g. beliefs, desires, intentions) and to predict others' behaviour based on those mental states. This ability forms an important basis for negotiating the large social groups and the complex social world in which humans live and have lived throughout their evolutionary history. Whether or not theory of mind is unique to humans among primates, evolution must build on structures that already exist, so theory of mind would have evolved by building on other social information-processing systems in the brains of our ancestors. Thus, whatever brain systems are involved in theory of mind

probably include those systems used for social information-processing in other primates (Brothers and Ring 1992).

Theory of mind involves both 'cold cognition'—inferences about others' *epistemic* states such as beliefs, knowledge, focus of attention—and 'hot cognition'—inferences about others' *affective* states such as emotions, preferences, beneficent or hostile intentions. Because it involves such a diverse set of high-level inferences, theory of mind is unlikely to be localized in a single brain region. Baron-Cohen and Ring (1994) have suggested that it would be more accurate to think of theory of mind inferences as being computed by a distributed neural circuit, with different regions contributing different types of computations. This chapter expands on their model, discussing several different brain regions that might play a role in theory of mind. At this point, there is only a small amount of empirical evidence relating to the brain basis of theory of mind, so any account must be speculative. This chapter reviews the evidence for the role of the amygdala, orbitofrontal cortex, medial frontal cortex, and dorsolateral frontal cortex. Focusing on the role that specific regions might play leaves out several important factors in a thorough account of the neural basis of theory of mind. This chapter does not explore processing differences between the right and left hemispheres; a thorough review of that literature is provided in the chapter by Brownell *et al.* in this volume. It also does not discuss the role of specific neurotransmitters, such as serotonin or oxytocin, that are thought to play a role in social behaviour (Raleigh and Brammer 1993; Raleigh *et al.* 1991; reviewed in Brothers 1994; Carter and Altemus 1997; Insel 1997; Nelson and Panksepp 1998). Furthermore, it does not discuss the role of particular patterns of connectivity between these areas. However, evidence for the role of the amygdala and frontal regions in theory of mind provides a starting point for further investigation into these other questions.

THE FRONTAL LOBES AND THEORY OF MIND

Because the frontal lobes seem to be involved in 'higher' cognition, it may be most profitable to look for structures that underlie social cognitive abilities, such as theory of mind, in frontal cortex. 'Prefrontal cortex', a term used to refer to cortex anterior to the premotor area, can be divided into somewhat functionally distinct subregions, with different consequences for damage to each area. There is agreement on two functionally important subregions.

1. dorsolateral frontal cortex, the upper (dorsal) and outer (lateral) surface of the frontal lobes including Brodmann's areas 6, lateral portions of 8–10, 44–46, regions that receive their blood supply from the middle cerebral artery; and
2. orbitofrontal cortex, primarily Brodmann's area 11, the ventral surface of the frontal lobes that sits above the eyes (the orbits) (Benson and Miller 1997; Bowen 1989; Kaczmarek 1984; Mattson and Levin, 1990). (See Fig. 11.1.)

Damage to dorsolateral frontal cortex produces a variety of general cognitive

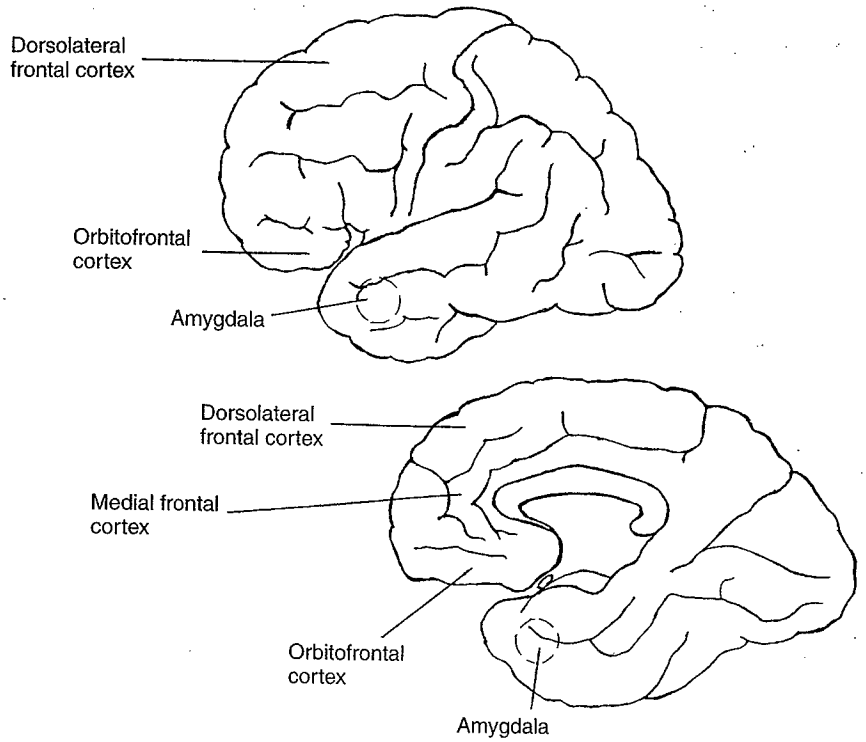

Fig. 11.1. Brain regions that may be involved in theory of mind. The upper diagram shows the outer surface of the left hemisphere. The lower diagram is of the medial surface of the right hemisphere, shown from a slice made between the two hemispheres. Only the approximate position of the amygdala is shown, as it is located inside the anterior temporal lobes, just in front of the hippocampus.

deficits, both low-level and high-level functions, such as novelty detection, inhibition of irrelevant sensory stimuli, temporal sequencing of events in memory, planning, executive function, task-switching, and working memory (Benson and Stuss 1986; Kimberg *et al.* 1997; Knight and Grabowecky 1995; Mattson and Levin 1990). Damage to orbitofrontal cortex, in contrast, produces primarily social and emotional changes: inappropriate humour, socially inappropriate behaviour (particularly verbal behaviour), self-centred behaviour, and a tendency to rambling, digressive speech (Alexander *et al.* 1989; Benson and Miller 1997; Bowen 1989; Damasio and VanHoesen 1983; Kaczmarek 1984; Mattson and Levin 1990).

There is less agreement on the functional importance of other subdivisions of prefrontal cortex. Medial frontal cortex is the inside surface of the frontal lobes between the two hemispheres, including Brodmann's areas 32, 12, and the medial portions of 8–10. Medial frontal cortex receives its blood supply from the anterior cerebral artery. Consequences of damage to this area may include apathy, akinesia, difficulties with language, and some inappropriate behaviour, depending on whether

the damage is more ventral or more dorsomedial (Alexander *et al.* 1989; Bowen 1989). Also, the term *dorso*lateral excludes areas of lateral frontal cortex that are not dorsal, such as area 47 and the lower parts of areas 10 and 45.

Orbitofrontal cortex and theory of mind

Each of the subregions of the frontal lobes may have its own contribution to make to theory of mind. Orbitofrontal cortex seems to be important in social cognition, because, of all the regions of the frontal lobes, damage there seems to most directly affect social behaviour. This discussion of the consequences of orbitofrontal damage includes patients with ventromedial cortex damage, that is, damage that includes both orbitofrontal cortex and the lower portion of medial frontal cortex. The two patient groups are discussed together because ventromedial damage includes orbitofrontal cortex damage, and because ventromedial frontal patients' social behaviour has been well characterized, and matches what has been described for patients with damage only to orbitofrontal cortex. Following Baron-Cohen and Ring (1994), this chapter argues that these problems in social behaviour arise from subtle impairments in theory of mind.

After damage to orbitofrontal or ventromedial cortex occurs, a person may go through a marked change in social behaviour. Patients with orbitofrontal or ventromedial damage often make inappropriate comments, particularly sexual comments, make inappropriate jokes, make poor choices in personal relationships, and have difficulty with the pragmatics of conversation (Alexander *et al.* 1989; Benson and Miller 1997; Damasio and VanHoesen 1983; Eslinger and Damasio 1985; Gronwall *et al.* 1998; Kaczmarek 1984; Mattson and Levin 1990; Saver and Damasio 1991). I have witnessed several examples of inappropriate behaviour while testing orbitofrontal patients. One patient opened the conversation when we went into a testing room by saying, 'OK, let's take our clothes off now.' Another patient, apropos of nothing we had been talking about, launched into a description of the pornographic novels he had been reading lately, and how he hoped that reading them would help save his marriage. (He and his wife were later separated.) None of these patients gave any indication that they thought they were saying anything inappropriate.

Though striking, such incidents do not occur during every social interaction. Difficulties with conversational pragmatics, however, are evident in any interaction with patients of this type. Orbitofrontal patients exhibit poor use of relevance, tending to drift from topic to topic, without giving the listener a clear sense of connection between topics (Alexander *et al.* 1989; Kaczmarek 1984). They do not appear to check whether the listener is interested in what they are saying. It is often difficult to get a word in edgewise, because they seem insensitive to signs that their conversational partner is trying to say something (Alexander *et al.* 1989). They will sometimes refer to something as if the listener knew what they were talking about, without having ever mentioned it; for example, a patient might say, 'It just wasn't the right job for him', without ever having mentioned what the job was. If the graduate student in the example at the beginning of the chapter had orbitofrontal damage, he probably would

have continued his conversation with the researcher when the president approached, insensitive to cues that the president was trying to enter the conversation.

Because of these social difficulties, patients with orbitofrontal or ventromedial damage often have difficulty maintaining friendships and intimate relationships. Furthermore, they are often taken advantage of by others. One patient that I tested told me about marrying a woman he had not known for very long. After the ceremony, she told him that she had married him only so that she could get a green card, and had no intention of having a relationship with him. He was genuinely surprised, and said that, up until that moment, he had no idea she was using him. Another patient had divorced his wife when his family discovered that she was stealing money from him. Eslinger and Damasio (1985) report that E.V.R., a patient with ventromedial damage, lost money by getting involved in several questionable business relationships.

The personality changes following orbitofrontal cortex damage have been described as 'pseudopsychopathic' or as 'acquired sociopathy' because of patients' relative lack of concern for how they affect others (Benson and Stuss 1986; Eslinger and Damasio 1985; Mattson and Levin 1990; Saver and Damasio 1991; Tranel 1994). However, these terms may be inappropriate. True psychopaths or sociopaths are characterized not only by their lack of concern for others, but also by their skill at manipulating or exploiting others. Patients with orbitofrontal damage do not have the necessary social skills to manipulate others; on the contrary, they are themselves vulnerable to being exploited. 'Socially impaired' would be a more accurate term. These patients' difficulty in reading the social information around them makes them appear unconcerned, but this difficulty also makes them vulnerable to others.

In contrast to their social skills, patients with damage to orbitofrontal cortex or ventromedial frontal cortex can (though do not always) perform normally on many tests of executive function and other cognitive functions, such as the Wisconsin Card Sorting Test, tests of verbal fluency, and some reasoning tasks (Mattson and Levin 1990). They are aware of their surroundings, know the place and time, and can be quite knowledgeable about current events. They may be somewhat distractible, but this appears primarily to affect preparation and execution of a response, rather than cognitive processing *per se* (Mattson and Levin 1990; Whyte *et al.* 1998). Thus, the pattern of symptoms following these types of damage is a pattern of notably impaired social cognition with general cognitive functioning relatively preserved.

Could this pattern reflect impairments in theory of mind? Theory of mind is certainly implicated in the pragmatics of language: the use of relevance and proper reference (Baron-Cohen 1995; Frith 1996; Happé 1994*a*,*b*; Tager-Flusberg 1993). Patients with damage to orbitofrontal cortex may not pay attention to what their conversational partner is interested in or knows about because they are not able to model others' minds easily. Theory of mind impairments could also result in inappropriate behaviour. Deciding whether or not to take a particular action involves having some model of what other people would think about it. From before their brain damage, these patients have a lifetime of experience knowing what is socially appropriate. Thus, they behave inappropriately not because they do not know from experience what is appropriate, but because they have difficulty computing 'on-line' during an interaction what would seem to others to be appropriate:

> Social rule-breaking is common . . . and it comes about *because the person no longer has the ability to judge how things [he/she] does are affecting other people.* (Gronwall *et al.* 1998, p. 65; emphasis added.)

Impairments in modelling others' mental states could make it difficult for someone to infer what others' reactions to certain actions would be.

If, as Baron-Cohen and Ring (1994) proposed, orbitofrontal cortex is part of a theory of mind circuit, then damage to orbitofrontal cortex would produce subtle impairments in theory of mind rather than the complete loss of mentalizing abilities. Orbitofrontal patients might be slower to make theory of mind inferences or might have difficulty integrating mental state inferences with other information. In real interactions, decisions have to be made rapidly and integrated with other information that is constantly changing, so real-time social interaction should pose the greatest challenge to a moderately impaired theory of mind mechanism.

Both neuroimaging studies and tests with lesion patients indicate that orbitofrontal cortex may be involved in theory of mind. Baron-Cohen *et al.* (1994), using SPECT, looked at areas that were activated during the recognition of mental state terms, such as 'remember', 'think', 'imagine'. Subjects listened to a list of words containing mental state terms and distractor terms matched for frequency, such as 'tape', 'district', or 'park', and were asked, for each word, to indicate whether it had to do with the mind or not. In the control condition, subjects listened to a different list of words, including words like 'shoulder' and 'teeth', as well as distractor terms, and were asked to indicate whether each word had to do with the body or not. They found that right orbitofrontal cortex relative to left frontal polar cortex was active during the mental state terms task.

Systematic tests of theory of mind in patients with orbitofrontal lesions have only been undertaken recently. If orbitofrontal cortex is part of a theory of mind circuit, then we would predict that patients with orbitofrontal cortex damage would not have difficulty on the most basic theory of mind tasks, but that it would take more difficult tasks to reveal their deficits. First order false-belief tasks can be passed by four-year-old children (Gopnik and Astington 1988; Wellman 1990; Wimmer and Perner 1983), second order false-belief tasks can be passed by five- to six-year-old children (Perner and Wimmer 1985), and adults are at ceiling on both types of tasks (Stone *et al.* 1998*a*). Thus, false-belief tasks test a very basic level of theory of mind inferences.

In order to test adults, we have devised a more developmentally advanced theory of mind task, a test of *faux pas* recognition (Stone *et al.* 1998). A *faux pas* occurs when someone says something awkward, hurtful, or insulting to another person, not knowing or not realizing that it should not have been said. For example, one of the items in the test gave the following story:

> Jeanette bought her friend Anne a crystal bowl for a wedding gift. Anne had a big wedding and there were a lot of presents to keep track of. About a year later, Jeanette was over one night at Anne's for dinner. Jeanette dropped a wine bottle by accident on the crystal bowl, and the bowl shattered. 'I'm really sorry, I've broken the bowl,' said Jeanette. 'Don't worry,' said Anne, 'I never liked it anyway. Someone gave it to me for my wedding.'

Recognizing a *faux pas* requires both an understanding of false or mistaken belief and an empathic inference about how the statement would affect someone. Seven-year-old children who could pass first and second order false-belief tasks did poorly on the *faux pas* task, but a larger proportion of children performed correctly on the task by age eleven; furthermore, twelve-year-olds with Asperger's syndrome were significantly impaired on the *faux pas* task, but not on compehension of control stories (Baron-Cohen *et al.* submitted). IQ was not correlated with performance on *faux pas* recognition in either the control subjects or the subjects with Asperger's syndrome.

Stone *et al.* (1998*a*) gave five patients with bilateral orbitofrontal cortex damage and five controls a series of theory of mind tasks that varied in difficulty: first-order false-belief tasks, second-order false-belief tasks, and the *faux pas* task. These patients scored at ceiling on the Mini Mental State Exam, a quick measure of basic cognitive functioning. The orbitofrontal patients performed well on first- and second-order false-belief tasks, indicating that their theory of mind abilities were intact at the seven-year-old level. They were impaired on the *faux pas* task, however (Stone *et al.* 1998*a*). (See Fig. 11.2.) Just as they often say inappropriate things in their own lives, they also have difficulty recognizing when someone says something inappropriate. Furthermore, we ran an additional 'empathy' condition where subjects were read the *faux pas* stories and asked 'How would [for example] Jeanette feel?' Even those patients who had failed to recognize *faux pas* had occurred could make the empathic inference that the character in the story would have felt hurt or insulted. However, they did not conclude from that fact that anything inappropriate had been said. Since we know from their performance on the false-belief tasks that their ability to represent 'cold' cognitive states like belief and knowledge is intact, and since they seem to have a basic level of empathic understanding, these results may indicate that they have difficulty combining information about mental states with affective information.

The *faux pas* task is a highly verbal task. However, theory of mind impairments in orbitofrontal patients are evident on less verbal tasks as well. We gave these same patients a series of face-processing tasks varying in difficulty that required them to infer someone else's internal state from information in the face. The more basic tasks tested whether they could recognize facial expressions of basic emotions (fear, anger, sadness, happiness, disgust, and surprise) from photographs of the full face. The more difficult task, 'Reading the Mind in the Eyes', tested their ability to read complex mental states from photographs of the eyes only (Baron-Cohen *et al.* 1997*b*). The orbitofrontal patients performed as well as or better than the controls did on recognizing basic facial experessions, but were significantly impaired on inferring mental states from information around the eyes (Stone *et al.* 1998*a*). (See Fig. 11.3.) A vocabulary test showed that they had full comprehension of the mental state terms used in this task, so their impairment cannot be accounted for by verbal comprehension difficulties.

These results from two different theory of mind tasks, one highly verbal and one more visual, indicate that orbitofrontal cortex may well be involved in theory of mind. In the future, how these patients and ventromedial patients perform on other advanced theory of mind tests should clarify the role of this region of the frontal lobes in theory of mind.

Alternative explanations of orbitofrontal patients' social deficits

Other theories have been proposed to account for the deficits in social behaviour that follow damage to orbitofrontal cortex. One theory that is commonly advanced to explain orbitofrontal patients' social deficits is that they are disinhibited. Thus, their inappropriate behaviour reflects what anyone would do without inhibition. Once a possibility for action comes to mind, these patients take the action, unable to inhibit the behaviour on the basis of possible consequences. While it may be true that orbitofrontal patients have some disinhibition, this is not a sufficient explanation for their social behaviour. Although they may say inappropriate things, they do not act on what they say. An orbitofrontal patient may make a verbal advance or lewd comment, but will not engage in lewd behaviour. If disinhibition is to be invoked as an explanation for what these patients say, then we are left without an explanation for what they do or rather, what they do not do. In addition, disinhibition does not provide a good account of their difficulties with conversational pragmatics. Inappropriate use of reference, failure to take account of common ground, and poor use of relevance do not follow in a straightforward way from disinhibition. Finally, it is not clear why being disinhibited would lead to failures of *faux pas* recognition or difficulties reading mental states from subtle expressions around the eyes. Thus, the hypothesis that orbitofrontal damage causes a partial impairment of the theory of mind mechanism provides a more concise explanation for the entire spectrum of social deficits exhibited by these patients than does the theory of disinhibition.

Another theory that has been proposed to account for the social and decision making deficits[1] of patients with ventromedial frontal damage is Damasio's 'Somatic Marker Hypothesis' (Damasio 1994). This theory postulates that the ventromedial region of the frontal lobes is critical for processing information about how emotional reactions are associated with particular imagined outcomes. These reactions are stored as memories of physiological emotional reactions, and when deciding which action to take, people rely on these 'gut feelings', or somatic markers, in choosing what to do.

> . . . *somatic markers are a special instance of feelings . . . [that] have been connected, by learning, to predicted future outcomes of certain scenarios* . . . Somatic markers . . . assist deliberation by highlighting some options (either dangerous or favorable) and eliminating them rapidly from subsequent consideration. (Damasio 1994, p. 174; emphasis in the original.)

People avoid bad outcomes because they have a physiological reaction to imagining the course of action leading to that outcome. Damasio hypothesizes that ventromedial cortex is crucial for processing information about these somatic markers. Ventromedial frontal patients have been found to lack autonomic responses to anticipated negative outcomes (Bechara *et al.* 1993). The somatic marker hypothesis can certainly account for inappropriate soical behaviour. Lacking the physiological response of imagined shame or embarrassment, a person would not have any reason to avoid inappropriate behaviour. This hypothesis could also possibly account for results on the *faux pas* test. Normally, upon hearing about another person's social gaffe, one experiences an emotional reaction of vicarious shame or embarrassment. It is possible

that, without this somatic response to reading about a *faux pas*, orbitofrontal patients cannot tell that something has been said that should not have been said. However, the 'Reading the Mind in the Eyes Task' does not depend on emotional responses in the same way, so the somatic marker hypothesis cannot account for both sets of results on theory of mind tests. It is also not clear how the somatic marker hypothesis could explain deficits in conversational pragmatics, such as inappropriate use of reference, and failure to take account of common ground.

Rolls (1996) has proposed that one of the major functions of orbitofrontal cortex is to represent changing reward values in the environment. There is strong evidence for this account, showing that orbitofrontal cortex is active in macaques when reversal learning or extinction occurs, and that earlier synapses in the temporal lobes are not involved (Rolls 1996). Studies of patients with orbitofrontal cortex lesions show that such patients have difficulty with reversal learning and extinction (Rolls *et al.* 1994). Rolls's theory by itself can explain some, but not all, of patients' inappropriate social behaviour and deficits in conversational pragmatics. However, this theory dovetails with the hypothesis that orbitofrontal cortex is involved in theory of mind. Orbitofrontal cortex is a large region, and is therefore likely to carry out several functions. It may be that in normal subjects, appropriate behaviour is maintained by punishment, such as disapproval, or by withholding of a reward, such as approval from the social environment. According to Rolls's theory, orbitofrontal patients would have difficulty learning from others' changing reactions to their behaviour. However, a theory of mind inference would be necessary to know whether another person approves or disapproves of one's behaviour. A combination of Rolls's theory and the 'theory of mind theory' could also explain some deficits in pragmatics. If a conversational partner initially seemed interested in a topic, and then lost interest, a patient with orbitofrontal damage would not be able to adjust behaviour based on this change in the social environment. A more cognitive explanation than Rolls's theory may be needed to account for other pragmatic deficits exhibited by patients with orbitofrontal damage. Relevance and common ground depend not on reward values in the environment, but on the semantic relatedness of different topics or on representing what someone else knows.

The fact that orbitofrontal cortex represents changing reward values also does not, by itself, account for the performance of patients with orbitofrontal damage on theory of mind tests. The *faux pas* task requires judgments about other people, and does not involve reward or punishment for one's own behaviour. Furthermore, in none of the *faux pas* stories is there information about a story character's reaction to the other character's *faux pas*, so there is no punishment or reward. The 'Reading the Mind in the Eyes Task' also does not involve any rewards or punishments, merely a choice between two mental state terms. The evidence is strong that orbitofrontal cortex does represent changing reward values. This account, combined with some theory of mind capacity in orbitofrontal cortex, can explain many of the social impairments of patients with orbitofrontal damage.

In short, while orbitofrontal patients may have some deficits in inhibition and in encoding 'somatic markers', neither of these theories gives a complete account of their social deficits. Rolls's theory provides a good account of how orbitofrontal damage disrupts affective responses. The theory that orbitofrontal damage also disrupts

theory of mind is able to explain the whole pattern of social problems suffered by orbitofrontal patients: inappropriate social behaviour, deficits in conversational pragmatics, and impairments on theory of mind tests. It is a parsimonious explanation for the social deficits (though not the nonsocial deficits) of patients with orbitofrontal cortex damage.

Medial frontal cortex and theory of mind

Just as with orbitofrontal patients, some patients with damage to medial frontal cortex may also exhibit social inappropriateness and problems with discourse. The literature on consequences of medial frontal damage leads to few firm conclusions, because the consequences of medial frontal damage can vary substantially, and correlations between type of symptom and location of lesion are not clear. Medial frontal damage can produce akinesia, particularly acutely, and difficulty initiating action (Alexander *et al.* 1989). Patients may say or do little, appearing apathetic and moving slowly, symptoms which may occur because of damage to the supplementary motor area (Alexander *et al.* 1989; Bowen 1989). Patients with medial frontal damage may also have some executive function deficits, showing impairment on the Wisconsin Card Sorting Test (Mattson and Levin 1990). Deficits in social behaviour and social cognition are not their most salient symptoms. However, patients with either right or left hemisphere lesions in medial frontal cortex have been noted to say inappropriate things or use humour inappropriately (Alexander *et al.* 1989). Left medial frontal lesions can also impair patients' ability to understand nonliteral language, such as metaphors and proverbs (Alexander *et al.* 1989). Benson and Stuss (1986) and Benson and Miller (1997) caution that there may not be a close mapping between exact symptoms and exact regions of the frontal lobes. Tentatively, one can say that socially inappropriate behaviour may result from medial frontal lesions that are more ventral; akinesia may result from more dorso-medial lesions, but the distinction between these two types of medial frontal damage is far from clear (Bowen 1989).

The socially inappropriate behaviour of medial frontal patients could be due to subtle deficits in theory of mind. Their pragmatic language difficulties could also arise from theory of mind deficits. Happé has noted that understanding the nonliteral use of language may often depend on theory of mind, so left medial frontal damage could well impair medial frontal patients' abilities in this domain because of a theory of mind deficit (Alexander *et al.* 1989; Happé 1994*a*,*b*).

Empirical support for the role of medial frontal cortex in theory of mind comes primarily from neuroimaging studies that have demonstrated left medial frontal activity during theory of mind tasks. Goel *et al.* (1995) found selective activation of left medial frontal cortex and left temporal cortex during a task requiring subjects to model another person's mental state about a target object. Control tasks required only a visual description of the object, memory retrieval or inferring the function of an object from its form. Fletcher *et al.* (1995) also found left medial frontal activation, in the medial portion of Brodmann's areas 8 and 9, during a task requiring mental state inferences compared with tasks requiring subtle physical inferences. They did not find any orbitofrontal cortex activation. Happé *et al.* (1996) further pinpointed left medial

frontal cortex, area 8, as important for theory of mind by showing that five individuals with Asperger's Syndrome did not show activity in area 8 during the same tasks.

The role of medial frontal cortex in theory of mind remains a puzzle to be sorted out. These neuroimaging studies found dorsal medial activation during theory of mind tasks, yet as noted above, it seems that social difficulties occur after medial damage that is more ventral than dorsomedial (Bowen 1989). Patients with medial frontal damage are relatively rare, because strokes are much more common in the middle cerebral artery than in the anterior cerebral artery, which supplies medial frontal cortex. Theory of mind has so far been tested in only one patient with medial frontal damage, an elderly patient with a prefrontal leucotomy whose damage included medial frontal regions (Bach *et al.* 1998). However, although the patient was impaired on second-order false-belief tests and advanced versions of Happé's Strange Stories, these deficits probably resulted from impairments in executive function, as the patient also had severe deficits on measures of executive function (Bach *et al.* 1998). No selective theory of mind deficits have yet been found in patients with selective medial frontal damage. In the next few years, systematic tests of theory of mind and executive function in patients with medial frontal damage should do much to illuminate these issues.

Dorsolateral frontal cortex and theory of mind

Dorsolateral frontal cortex may also be involved in theory of mind, if only because the general cognitive functions of dorsolateral frontal cortex are necessary in carrying out some theory of mind computations. False-belief tasks, for example, place strong demands on inhibitory control, sequencing and working memory (Ozonoff *et al.* 1991; Pennington *et al.* 1997; Stone *et al.* 1998*a*). To solve a first- or second-order false-belief task, a subject must keep all the elements of the story in working memory before the questions are asked—the original location of the object, where the object was moved, where each character was when object was moved—and must remember them in the proper sequence. Furthermore, answering the 'belief' question correctly requires inhibiting what the subject knows to be true in order to answer with what one of the story characters believes to be true. Reality may be more salient than a story character's beliefs, and executive control may be required to inhibit responses based on the real location of the object.

When Stone *et al.* (1998*a*) tested orbitofrontal patients on theory of mind tasks, five patients with left dorsolateral frontal damage were also tested on the same tasks. We manipulated the working memory demands of the false-belief tasks by running them in two conditions: with and without a memory load. For example, one of the stories was about Tony, who came into the kitchen and put a bottle of Coke in a cabinet. After he left the room, a woman named Maria came in and moved the Coke into the refrigerator. Later, Tony came back in. In the 'memory load' condition, subjects watched the action of this story on videotape as the experimenter told the story, and had to keep the elements of the story in working memory in proper sequence to answer the questions. For the 'no memory load' condition, we printed out video stills of the action in the story, and laid them out in front of the subjects as the experimenter told the story. All the pictures were visible while subjects answered the

questions. Thus, the subjects did not have to remember the story elements or sequence them. This manipulation had no effect on orbitofrontal patients' performance.

Dorsolateral frontal patients, in contrast, performed much better in the 'no memory load' condition (see Fig. 11.2). Working memory load significantly impaired their performance on false-belief tasks. However, even when they made errors in the 'memory load' condition, they were not more likely to make errors on the belief questions as opposed to the other questions (Stone *et al.* 1998*a*). Thus their performance overall gave us no reason to conclude that they have any deficits in making mentalistic inferences *per se*. On the *faux pas* task, they made errors only when they got confused about the details of the stories, even though the stories were always right in front of them (Stone *et al.* 1998*a*). These patients with dorsolateral frontal damage had lower scores than the orbitofrontal patients on the Mini Mental State Exam. Thus, for these two groups, performance on this test of general cognitive functioning does not predict performance on the *faux pas* task.

One weakness of this study is that it compared patients with bilateral orbitofrontal damage with patients with unilateral dorsolateral frontal damage. We can conclude from the results that left dorsolateral cortex alone does not seem to be critical for making mental state inferences. However, it is still possible that bilateral damage to dorsolateral frontal cortex would produce theory of mind deficits. Price *et al.* (1990) report two adult patients with bilateral dorsolateral frontal damage acquired early in life. They tested these patients on a perspective-taking task, which does require a

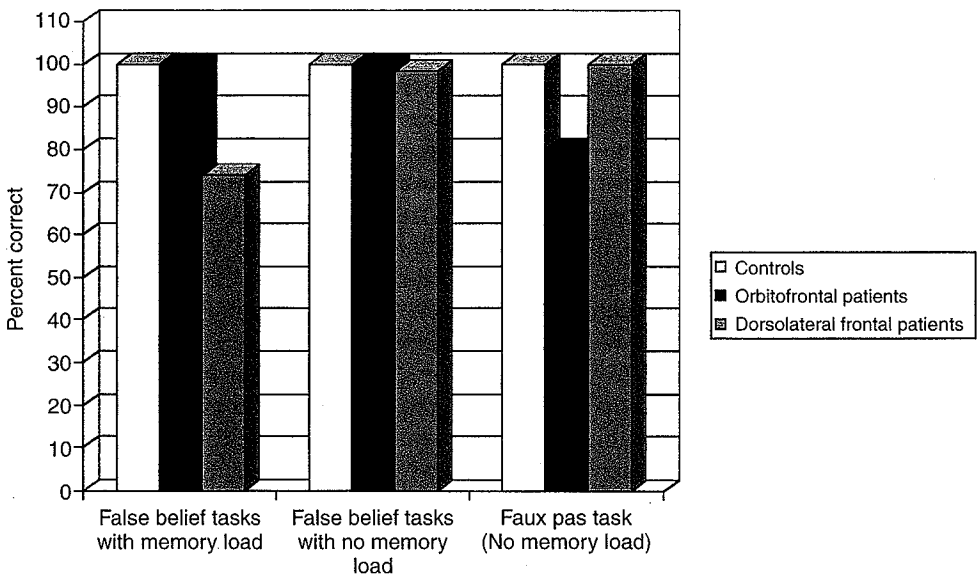

Fig. 11.2. Performance of frontal patients and controls on theory of mind tasks. N = 5 in each group of subjects. Because there were no strong differences in performance between first- and second-order false-belief tasks in any groups, performance on false-belief tasks in this graph was averaged over first- and second-order tasks.

theory of mind. The subject looked at a map of a town and was told that someone at a certain location on the map was lost and needed to get to a particular house on the map. The experimenter then read a set of directions for getting from where the lost person was to the house. The directions contained four different ambiguities such that a person could make mistakes and end up at the wrong house. After reading the directions, the subjects were asked to say which parts of the directions were ambiguous and could have led the lost person to make a mistake. These two patients failed this task.

The task could be seen as a theory of mind task, because it depends on understanding false-belief. However, it also places demands on working memory, because the subject has to keep all of the directions in memory to answer the question. Thus, these patients could have failed on this task because of working memory limitations rather than because their theory of mind was impaired. Further studies, investigating how bilateral dorsolateral frontal patients perform on theory of mind tasks that are controlled for executive function demands, are needed.

THE AMYGDALA AND THEORY OF MIND

> The finely tuned links between the meanings of social stimuli on the one hand, and the patterns of physiologic activity set into play on the other, presumably are embodied in the network of intrinsic connections in the amygdala. (Kling and Brothers 1992, p. 356.)

The amygdala is a small almond-shaped structure that is located in the medial temporal lobes, just anterior to the hippocampus. It receives polysensory connections from many cortical areas, and sends projections to the hypothalamus, which serve to initiate autonomic responses, ventro–striatal areas, which provide a pathway to initiate motor responses, and to temporal, orbitofrontal, and insular cortex, which may provide a cognitive output (Everitt and Robbins 1992; Halgren 1992; Rolls 1992). The amygdala serves several general functions within the domain of emotion and social behaviour. While it cannot be specific to theory of mind, it forms an important input system to the theory of mind circuit.

The amygdala processes emotional responses and information about reward (Aggleton 1992; Gaffan 1992; Halgren 1992; Kling and Brothers 1992; LaBar and LeDoux 1997; Rolls 1992). In particular, the lateral and central nuclei of the amygdala have been implicated in fear responses (LeDoux 1988, 1990). Natural selection has designed us so that those activities that enhance fitness are rewarding, and elicit positive emotions to draw us towards those activities, while those that are deleterious to fitness elicit negative emotions to push us away from those activities. The amygdala, therefore, should be involved in any highly fitness-relevant situation. It will be activated in recognizing the situation, and in producing the adaptive response, for example, recognizing a dangerous predator and initiating a fear response: freezing or flight (Kling and Brothers 1992; LeDoux *et al.* 1988, 1990). The set of stimuli that activate the amygdala should vary from species to species, as what is adaptive also varies.

Many types of information in the social environment are fitness-relevant. Direct eye contact signifies that one is the object of a conspecific's attention. It may be the first signal in an attack, a mating attempt, or an initiation of social exchange, any of which could affect one's fitness. Being the object of another's attention is unlikely to be fitness-neutral. Accordingly, people respond to direct eye contact autonomically, and even young infants respond strongly to eye contact (Baron-Cohen 1995). Young *et al.* (1995) note that the amygdala is strongly interconnected with the superior temporal sulcus, which, in macaques, has been found to be critical to determining gaze direction. In human subjects, bilateral amygdalotomy impairs the ability to judge gaze direction, and in particular, to judge direct eye contact (Young *et al.* 1995). It is unknown whether this impairment results from a geometric problem computing gaze direction or from a lack of the usual emotional response to eye contact. However, we can speculate that recognizing and responding affectively to the socially significant state 'looking at me' may be one of the important soical functions of the amygdala.

Information about others' eye gaze direction is a central building block for theory of mind. Joint attention, understanding reference, and knowing what someone else knows based on what they have seen, all depend on information about gaze direction (Baron-Cohen 1995; Baron-Cohen *et al.* 1997*a*). Thus, if the amygdala is involved in marking information about others' gaze direction as significant, then the amygdala must form a crucial input to the theory of mind mechanism.

Because others' emotional states can have a significant impact, information about others' emotions is also fitness-relevant. The amygdala appears to be involved in recognizing emotional displays. In monkeys, cells in the amygdala can be selectively activated in response to facial expressions of threat, predator warning vocalizations, infants' vocalizations on being separated from the mother, and displays of a conspecific approaching (Brothers and Ring 1992; Kling and Brothers 1992). In humans, fMRI studies have shown that the amygdala is active during presentation of the facial expression of fear (Morris *et al.* 1996; Phillips *et al.* 1997). People who have suffered bilateral amygdala damage have difficulty recognizing facial and vocal expressions of emotion, particularly fear and anger (Adolphs *et al.* 1994; Broks *et al.* 1998; Calder *et al.* 1996; Scott *et al.* 1997; Young *et al.* 1995). Of course, the significance of another's affect may depend on who he or she is, so information about identity can also be important in determining the response to another's affective signal. Cells in the amygdala have been found to respond to the individual identity of conspecifics (Kling and Brothers 1992; Rolls 1992). Information about others' affective mental states may also be an important input to the theory of mind mechanism.

Halgren (1992) notes that, in humans, the amygdala receives information from the cortex that is already highly processed, and can generate an emotional response not just to concrete percepts, but also to words, thoughts, mental images, and 'other meaningful stimuli that have previously been associated with visceral upset' (p. 213). Thus, even complex situations or abstract ideas (such as a belief about another person's mental state) may cause emotional reactions. The graduate student in the example at the beginning of the chapter could have an emotional reaction, generated in the amygdala, to the thought, 'Maybe that researcher was relieved that the president came up, because she really didn't want to talk to me.' Thus, the amygdala

may be interwoven with the other parts of the theory of mind mechanism, involved in generating emotional reactions on the basis of others' mental states. Consistent with this, stimulation of the amygdala may produce complex social emotions: the impression of being criticized, socially isolated, or threatened, or the impression that another person is demanding (Brothers 1994; Brothers and Ring 1992; Kling and Brothers 1992).

There is some empirical support for a role for the amygdala in theory of mind. Baron-Cohen *et al.* (1999) have found amygdala activation in normal subjects during an fMRI scan while subjects are doing the 'Reading the Mind in the Eyes Task', inferring complex mental states from images of the eyes, relative to a control task. Furthermore, they found that the amygdala was not active during this task when they scanned individuals with Asperger's Syndrome. Stone *et al.* (1998*b*) also found that two patients with bilateral amygdala damage were impaired on this task, relative to controls. (See Fig. 11.3.) In addition, one bilateral amygdalotomy patient was significantly impaired on the *faux pas* test (Stone *et al.* 1998*b*).

Baumann and Kemper (1985) reported that the brains of autistic subjects studied on autopsy showed that cells in the amygdala were unusually densely packed. If the amygdala is a critical input system to the theory of mind mechanism, and that input system is not functioning properly in individuals with autism because of abnormal cell growth patterns, this could be an important factor in their abnormal development of a theory of mind.

CONCLUSIONS

Scholars have started collecting direct empirical evidence on the role of each of these brain regions in theory of mind only in the past few years. Based on the curently available evidence, there is reason to conclude that orbitofrontal cortex, medial frontal cortex, dorsolateral frontal cortex, and the amygdala are all involved in theory of mind computations. Research in coming years can provide information about the

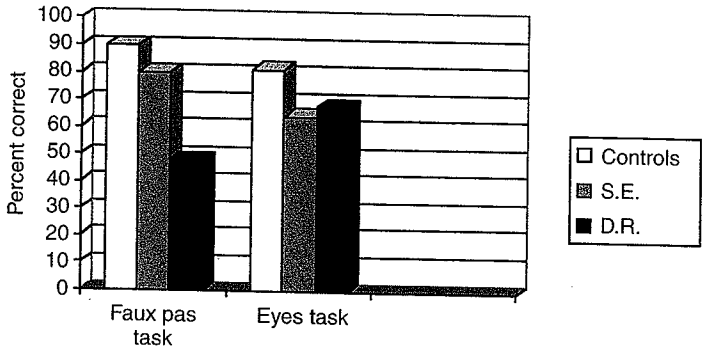

Fig. 11.3. Performance of two patients with bilateral amygdala damage, S. E. and D. R., on two theory of mind tasks.

precise role of each region. Brothers and Ring (1992) have talked about the distinction between 'cold' and 'hot' aspects of theory of mind, that is, between representations of the purely cognitive states of others, such as belief and knowledge, and representations of their affective and motivational states, such as emotions or hostile intentions. One possible area of inquiry is the role of each region in 'hot' and 'cold' mental-state inferences. The amygdala seems to form an important input to one of the earliest-developing and purely cognitive aspects of theory of mind, joint attention, by getting the developing brain to pay attention to others' eye gaze. It may be productive to investigate the role of the amygdala specifically in joint attention. Another important question is the nature of the contribution of each cerebral hemisphere. Fletcher *et al.* (1995) and Goel *et al.*, (1995) found selective left medial frontal activation during theory of mind tasks, whereas Happé *et al.* (1999) and Winner *et al.* (1998) have found theory of mind deficits in patients with right hemisphere damage. These results are not inconsistent, but the exact role of each hemisphere can be clarified in the future. Stone *et al.* (1998*a*) and Stone *et al.* (1998*b*) found theory of mind deficits in patients with bilateral orbitofrontal damage, but not in patients with unilateral left dorsolateral frontal damage. Price *et al.* (1990) also found a possible theory of mind deficit in patients with bilateral dorsolateral frontal damage. The deficits in theory of mind and in recognition of gaze direction that have been found in amygdala patients have been reported for patients with bilateral amygdala damage (Stone *et al.* 1998*b*; Young *et al.* 1995). More study is needed to determine if bilateral damage to particular brain structures is necessary to disrupt theory of mind.

Neuroimaging and lesion studies have different and complementary contributions to make to future research into theory of mind. Neuroimaging studies can reveal whether an area is involved in a particular task, but not whether that area is critical for that task. Studies with lesion patients can supplement neuroimaging studies by revealing whether a particular region is critical for a particular function. However, because lesion location varies within patient groups and because often a large area with more than one structure is damaged in lesion patients, it can be difficult to get a high degree of anatomical specificity from lesion studies. Neuroimaging studies may provide a more precise picture of the exact regions involved. The two methods should work in concert in future research.

Any account of the neural basis of theory of mind written at this point must be speculative. In the years that come, systematic tests that tap into all of these different aspects of theory of mind can be used with focal lesion patients and neuroimaging paradigms to give us a clearer picture of the contribution each brain region makes to this vital social cognitive ability.

Ackowledgements

Preparation of this chapter was supported by grants from Trinity College at the University of Cambridge, the Solomon R. and Rebecca D. Baker Foundation, and the Cure Autism Now Foundation. I am very grateful for the opportunity I had to spend time in the Department of Experimental Psychology at the University of Cambridge as a visiting scholar in 1997–1998 when I began writing this chapter. I

thank Simon Baron-Cohen, Bruce Pennington, Melissa Rutherford, and Piotr Winkielman for thoughtful comments on early drafts of this chapter. Finally, without the rich intellectual environment provided by discussions with my collaborators and colleagues, Andy Calder, Jill Keane, Bob Knight, Michele O'Riordan, Kate Plaisted, and Andy Young, the ideas presented here could not have taken shape. Special thanks go to Simon Baron-Cohen for his tremendous colleagueship over the past five years. His collaboration has been essential and invaluable in everything presented in this chapter.

Note

1. Both ventromedial patients and patients with pure orbitofrontal damage may show inappropriate social behaviour. In addition, Damasio describes ventromedial patients as going through endless deliberations to make even simple decisions, such as when to schedule an appointment or what to wear. The patients I have examined with only orbitofrontal damage and no medial frontal damage do not have these difficulties with decision making. Thus, these difficulties in ventromedial patients may be caused specifically by medial frontal damage, and could have to do with deficits in response selection, rather than a lack of somatic markers.

REFERENCES

Adolphs, R., Tranel, D., Damasio, H. and Damasio, A. R. (1994). Impaired recogniton of emotion in facial expressions following bilateral damage to the human amygdala. *Nature*, **372**, 669–72.

Aggleton, J. P. (1992). The functional effects of amygdala lesions in humans: a comparison with findings from monkeys. In *The amygdala: neurobiological aspects of emotion, memory and mental dysfucntion* (ed. J. P. Aggleton), pp. 485–503. Wiley-Liss, New York.

Alexander, M. P., Benson, D. F. and Stuss, D. T. (1989). Frontal lobes and language. *Brain and Language*, **37**, 656–91.

Bach, L., Davies, S., Colvin, C., Wijeratne, C., Happé, F. and Howard, R. (1998). A neuropsychological investigation of theory of mind in an elderly lady with frontal leucotomy. *Cognitive Neuropsychiatry*, **3**(2), 139–59.

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

Baron-Cohen, S. and Ring, H. (1994). A model of the mindreading system: neuropsychological and neurobiological perspectives. In *Origins of an understanding of mind* (ed. P. Mitchell and C. Lewis). Erlbaum, Hillsdale, NJ.

Baron-Cohen, S., Ring, H., Moriarty, J., Shmitz, P., Costa, D. and Ell, P. (1994). Recognition of mental state terms: clinical findings in children with autism and a functional neuroimaging study of normal adults. *British Journal of Psychiatry*, **165**, 640–9.

Baron-Cohen, S., Ring, H., Wheelwright, S., Bullmore, E., Brammer, M., Simmons, A. and Williams. S. (1999). Social intelligence in the normal and autistic brain: an fMRI study. *European Journal of Neuroscience*, **11**, 1891–8.

Baron-Cohen, S., Baldwin, D. A. and Crowson, M. (1997*a*). Do children with autism use the

speaker's direction of gaze strategy to crack the code of language? *Child Development*, **68**(1), 48–57.

Baron-Cohen, S., Joliffe, T., Mortimore, C. and Robertson, M. (1997*b*). Another advanced test of the theory of mind: evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry*, **38**(7), 813–22.

Baron-Cohen, S., O'Riordan, M., Stone, V. E., Jones, R. and Plaisted, K. (in press). Recognition of faux pas by normally developing children and children with Asperger Syndrome. *Journal of Autism and Developmental Disorders.*

Baumann, M. L. and Kemper, T. (1985). Histoanatomic observations of the brain in early infantile autism. *Neurology*, **35**, 866–74.

Bechara, A., Tranel, D., Damasio, H. and Damasio, A. R. (1993). Failure to respond autonomically in anticipation of future outcomes following damage to human prefrontal cortex. *Society for Neuroscience Abstracts*, **19**, 791.

Benson, D. F. and Miller, B. L. (1997). Frontal lobes: clinical and anatomic aspects. In *Behavioral neurology and neuropsychology* (ed. T. E. Feinberg and M. J. Farah), pp. 401–8. McGraw Hill, New York.

Benson, D. F. and Stuss, D. T. (1986) *The frontal lobes.* Raven Press, New York.

Bowen, M. (1989). Frontal lobe function. *Brain Injury*, **3**, 109–28.

Broks, P., Young, A. W., Maratos, E. J., Coffey, P. J., Calder, A. J., Isaac, C. L., Mayes, A. R., Hodges, J. R., Montaldi, D., Cezayirli, E., Roberts, N. and Hadley, D. (1998). Face processing impairments after encephalitis: amygdala damage and recognition of fear. *Neuropsychologia*, **36**(1), 59–70.

Brothers, L. (1997). *Friday's footprint: how society shapes the human mind.* Oxford University Press, New York.

Brothers, L. and Ring, B. (1992). A neuroethological framework for the representation of minds. *Journal of Cognitive Neuroscience*, **4**(2), 107–18.

Calder, A. J., Young, A. W., Rowland, D., Perrett, D. I., Hodges, J. R. and Etcoff, N. L. (1996). Facial emotion recognition after bilateral amygdala damage: differentially severe impairment of fear. *Cognitive Neuropsychology*, **13**(5), 699–745.

Carter, C. S. and Altemus, M. (1997). Integrative functions of lactational hormones in social behavior and stress management. *Annals of the New York Academy of Science*, **807**, 164–74.

Damasio, A. R. (1994). *Descartes' error: emotion, reason and the human brain.* Avon Books, New York.

Damasio, A. R. and VanHoesen, G. W. (1983). Emotional disorders associated with focal lesions of the limbic frontal lobe. In *Neuropsychology of human emotion* (ed. K. M. Heilman and P. Satz), pp. 85–110. Guilford Press, New York.

Eslinger, P. J. and Damasio, A. R. (1985). Severe disturbance of higher cognition after bilateral frontal lobe ablation: patient EVR. *Neurology*, **35**, 1731–41.

Everitt, B. J. and Robbins, T. W. (1992). Amygdala-ventral striatal interactions and reward-related processes. In *The amygdala: neurobiological aspects of emotion, memory and mental dysfunction* (ed. J. P. Aggleton), pp. 401–29. Wiley–Liss, New York.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J. and Frith, C. (1995). Other minds in the brain: a functional imaging study of 'theory of mind' in story comprehension. *Cognition*, **57**, 109–28.

Frith, C. (1996). Brain mechanisms for 'having a theory of mind'. *Journal of Psychopharmacology*, **10**(1), 9–15.

Gaffan, D. (1992). Amygdala and the memory of reward. In *The amygdala: neurobiological aspects of emotion, memory and mental dysfunction* (ed. J. P. Aggleton), pp. 471–83. Wiley–Liss, New York.

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

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

Gronwall, D., Wrightson, P. and Waddel, P. (1998). *Head injury: the facts.* Oxford University Press.

Halgren, E. (1992). Electrophysiological responses in the human amygdala. In *The amygdala: neurobiological aspects of emotion, memory and mental dysfunction* (ed. J. P. Aggleton), pp. 191–228. Wiley–Liss, New York.

Happé, F. (1994*a*). 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.

Happé, F. (1994*b*). Communicative competence and theory of mind in autism: a test of Relevance Theory. *Cognition*, **48**, 101–19.

Happé, F., Ehlers, S., Fletcher, P. C., Frith, U., Johansson, M., Gillberg, C., Dolan, R., Frackowiak, R. and Frith, C. (1996). 'Theory of mind' in the brain: evidence from a PET scan study of Asperger syndrome. *Neuroreport*, **8**(1), 197–201.

Happé, F., Brownell, H. and Winner, E. (1999). Acquired theory of mind impairments following right hemisphere stroke. *Cognition*, **70**, 211–40.

Insel, T. R. (1997). A neurobiological basis of social attachment. *American Journal of Psychiatry*, **154**(6), 726–35.

Kaczmarek, B. L. J. (1984). Neurolinguistic analysis of verbal utterances in patients with focal lesions of frontal lobes. *Brain and Language*, **21**, 52–8.

Kimberg, D. Y., D'Esposito, M. and Farah, M. J. (1997). Frontal lobes: cognitive neuropsychological aspects. In *Behavioral neurology and neuropsychology* (ed. T. E. Feinberg and M. J. Farah), pp. 409–18. McGraw-Hill, New York.

Kling A. S. and Brothers, L. A. (1992). The amygdala and social behavior. In *The amygdala: neurobiological aspects of emotion, memory and mental dysfunction* (ed. J. P. Aggleton), pp. 353–77. Wiley–Liss, New York.

Knight, R. T. and Grabowecky, M. (1995). Escape from linear time: prefrontal cortex and conscious experience. In *The cognitive neurosciences* (ed. M. S. Gazzaniga). MIT Press, Cambridge, MA.

LaBar, K. S. and LeDoux, J. E. (1997). Emotion and the brain: an overview. In *Behavioral neurology and neuropsychology* (ed. T. E. Feinberg and M. J. Farah), pp. 675–89. McGraw-Hill, New York.

LeDoux, J. E., Iwata, J., Cicchetti, P. and Reis, D. J. (1988). Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *Journal of Neuroscience*, **8**, 2517–29.

LeDoux, J. E., Cicchetti, P., Xagoraris, A. and Romanski, L. M. (1990). The lateral amygdaloid nucleus: sensory interface of the amygdala in fear conditioning. *Journal of Neuroscience*, **10**, 1062–9.

Mattson, A. J. and Levin, H. S. (1990). Frontal lobe dysfunction following closed head injury. *Journal of Nervous and Mental Disease*, **178**(5), 282–91.

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder, A. J. and Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature*, **383**, 812–5.

Nelson, E. E. and Panksepp, J. (1998). Brain substrates of infant-mother attachement:

contributions of opioids, oxytocin and norepinephrine. *Neuroscience and Biobehavioral Research*, **22**(3), 437–52.

Ozonoff, S., Rogers, S. J. and Pennington, B. F. (1991). Executive function deficits in high-functioning autistic individuals: relationship to theory of mind. *Journal of Child Psychology and Psychiatry*, **32**, 1107–22.

Pennington, B. F., Rogers, S. J., Bennetto, L., Griffin, E. M., Reed, D. T. and Shyu, V. (1997). Validity tests of the executive dysfunction hypothesis of autism. In *Autism as an executive disorder* (ed. J. Russell), pp. 143–73. Oxford University Press.

Perner, J. and Wimmer, H. (1985). 'John thinks that Mary think that . . .': attribution of second-order false beliefs by five to ten year old children. *Journal of Experimental Child Psychology*, **39**, 437–71.

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrews, C., Calder, A. J., Bullmore, E. T., Perrett, D. I., Rowland, D., Williams, S. C. R., Gray, J. A. and David, A. S. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature*, **389**, 495–8.

Price, B., Daffner, K., Stowe, R. and Mesulam, M. (1990). The comportmental learning disabilities of early frontal lobe damage. *Brain*, **113**, 1383–93.

Raleigh, M. J. and Brammer, G. L. (1993). Individual differences in serotonin-2 receptors in social behavior in monkeys. *Society for Neuroscience Abstracts*, **19** 592.

Raleigh, M. J., McGuire, M., Brammer, G. L., Pollack, D. and Yuwiler, A. (1991). Serotonergic mechanisms promote dominance acquisition in adult male vervet monkeys. *Brain Research*, **559**, 181–90.

Rolls, E. T. (1992). Neurophysiology and functions of the primate amygdala. In *The amygdala: neurobiological aspects of emotion, memory and mental dysfunction* (ed. J. P. Aggleton), pp. 143–65. Wiley–Liss, New York.

Rolls, E. T. (1996). The orbitofrontal cortex. *Philosophical Transactions of the Royal Society*, **351**, 1433–43.

Rolls, E. T., Hornak, J., Wade, D. and McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery and Psychiatry*, **57**(12), 1518–24.

Saver, J. L. and Damasio, A. R. (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. *Neuropsychologia*, **29**, 1241–9.

Scott, S. K., Young, A. W., Calder, A. J. and Hellawell, D. J. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature*, **385**, 254–7.

Stone, V. E., Baron-Cohen, S. and Knight, R. T. (1998*a*). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, **10**(5), 640–56.

Stone, V. E., Baron-Cohen, S., Young, A. W., Calder, A. and Keane, J. (1998*b*). Impairments in social cognition following orbitofrontal or amygdala damage. *Society for Neuroscience Abstracts*, **24**, 1176.

Tager-Flusberg, H. (1993). What language reveals about the understanding of minds in children with autism. In *Understanding other minds: perspectives from autism* (1st Edn) (ed. S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen). Oxford University Press.

Tranel, D. (1994). 'Acquired sociopathy': the development of sociopathic behavior following focal brain damage. *Progress in Experimental Personality and Psychopathology Research*, 285–311.

Wellman, H. M. (1990). *The Child's Theory of Mind*. MiT Press, Cambridge, MA.

Whyte, J., Fleming, M., Cavallucci, C. and Coslett, H. B. (1998). The effects of visual distraction following traumatic brain injury. *Journal of the International Neuropsychological Society*, **4**(2), 127–36.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, **13**, 103–28.

Winner, E., Brownell, H., Happé, F., Blum, A. and Pincus, D. (1998). Distinguishing lies from jokes: theory of mind deficits and discourse interpretation in right hemisphere brain-damaged patients. *Brain and Language*, **62**, 89–106.

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

UNDERSTANDING OTHER MINDS

Perspectives from Developmental Cognitive Neuroscience

Second Edition

Edited by

SIMON BARON-COHEN
Lecturer in Psychopathology,
University of Cambridge

HELEN TAGER-FLUSBERG
Professor of Psychology,
University of Massachusetts

and

DONALD J. COHEN
Professor of Child Psychiatry, Pediatrics, and Psychology,
Director, Child Study Center,
Yale University

OXFORD
UNIVERSITY PRESS