Understanding orbitofrontal contributions to theory-of-mind reasoning: Implications for autism

Mark A. Sabbagh

Department of Psychology, Queen's University, Kingston, Canada K7L 3N6

Accepted 17 April 2003

Abstract

Autism is a lifelong developmental disorder that is associated with severe difficulties with “theory-of-mind”—the understanding that others’ behaviors are motivated by internal mental states. Here, we raise the possibility that research examining the neural bases of theory-of-mind reasoning has the potential to inform researchers about the elusive functional neural impairments associated with autism. Evidence from our lab and others’ suggests that theory-of-mind reasoning may be fractionated into at least two functionally and anatomically distinct neural circuits. Specifically, the ability to decode others’ mental states from observable cues (such as facial expressions) may rely on contributions from the orbitofrontal/medial temporal circuit within the right hemisphere. In contrast, the ability to reason about others’ mental states may rely on left medial frontal regions. We conclude by reviewing evidence suggesting that the developmental roots of autism might lie in abnormal functioning of the orbitofrontal/medial temporal circuit which may, in turn, underlie the abnormal development of social-cognitive skills among individuals with autism.

1. Introduction

Researchers generally agree that human social interaction gains much of its richness and complexity because we possess a “theory-of-mind”—an understanding of how other’s behaviors are motivated by their internal mental states, such as beliefs, desires, intentions, and emotions (Wellman, 1990). The conceptual framework underlying research on theory of mind has its roots in developmental psychology. At present, issues surrounding aspects of children’s development of a theory-of-mind have come to guide research on children’s social-cognitive development (Taylor, 1996). Through this work, researchers have established links between theory-of-mind and executive functions (Carlson & Moses, 2001; Frye, Zelazo, & Burack, 1998; Perner & Lang, 1999), language development (Baldwin & Tomasello, 1998; Bloom, 2000; Sabbagh & Baldwin, 2001), prosocial behavior, and emotional functioning (Dunn, 1999; Happé & Frith, 1996).

Perhaps because of its wide ranging applications, theory-of-mind research has extended well beyond the domain of developmental psychology. Indeed, the topic has become an interdisciplinary hub for clinical, social, cognitive, and comparative psychologists and philosophers interested in understanding the foundations of everyday social cognition (Malle, Moses, & Baldwin, 2001). Still more recently, cognitive neurologists and neuroscientists have begun to explore the neural bases of theory-of-mind skills (see Siegal & Varley, 2002, for a recent review). This latter branch of research is contributing to our developing understanding of how the cognitive skills that enable high-level social understanding are organized in the human brain (Cacioppo et al., 2002). Perhaps most importantly, this research has opened new windows for understanding the neuropathological bases of clinical disorders in which social-cognitive and theory-of-mind skills may be specifically impaired (e.g., Doody, Goetz, Johnstone, Frith, & Cunningham-Owens, 1998).

One disorder in which a theory-of-mind approach has been especially fruitful is autism (Baron-Cohen, 1995). Autism is a lifelong and debilitating neuro-cognitive developmental disorder that affects between 1/250
(Gillberg & Wing, 1999) and 1/500 (Fombonne, 1999) individuals. Autism is characterized by a triad of features, including impaired social development, delayed and deviant language, and an insistence on sameness (American Psychiatric Association, 1994; Rutter, 1978). Baron-Cohen (Baron-Cohen, 1995; Baron-Cohen, Leslie, & Frith, 1985) was one of the first to suggest that this apparently disparate triad of symptoms may stem from a fundamental deficit in theory-of-mind reasoning. Through his pioneering and continuing studies, it is now well-established that many individuals with autism, including high-functioning autism and Asperger syndrome, experience difficulties engaging theory-of-mind reasoning (see Baron-Cohen, 2001, for a review).

What is perhaps most compelling about many of these cases is that, especially among high-functioning individuals with autism, theory-of-mind deficits stand in contrast to surprising strengths in other areas of cognitive functioning. For instance, individuals with autism have difficulty recognizing emotional mental states in pictures of eyes, but no trouble discerning other underlying traits (such as gender or age) (Baron-Cohen, Wheelwright, & Joliffe, 1999). Also, individuals with autism have difficulty reasoning about others’ false beliefs (i.e., beliefs that do not match the true state of the world), but are adept at reasoning about non-mental representations that also can be false such as photographs (Leslie & Thaiss, 1992). This pattern of findings has been taken by some to suggest that individuals with autism’ theory-of-mind deficit represents a rather specific neuro-cognitive impairment (Baron-Cohen, 1994).

While there is general agreement now that social-cognition is particularly impaired in autism, there is considerable debate regarding the best way to characterize the neuro-cognitive nature of this impairment (see commentary on Baron-Cohen, 1994). In our lab, we have been attempting to gain insight into this question by characterizing the neural correlates of theory-of-mind reasoning in normally developing adults using the same tasks that have been used to index a theory of mind impairment in autism. This work has two main research aims. First, by gaining an appreciation of the functional cortical regions that are important for theory-of-mind reasoning, we can begin to develop hypotheses regarding the neuro-cognitive processes that distinguish theory-of-mind reasoning from reasoning in other domains. Second, insights from this approach can, in turn, provide a way of understanding how known neuro-cognitive characteristics associated with autism may map onto the primary social-cognitive deficits that characterize this developmental disorder.

1.1. Separating component processes in theory-of-mind

Before discussing our work, it is important to functionally define what is meant by Theory-of-Mind in a way that makes it amenable to cognitive neuroscience investigations. Researchers typically functionally define theory-of-mind as a unitary skill that is engaged anytime one needs to represent the mental states of others. Sometimes, the skill is referred to by the term mind reading (Whiten, 1991). Nonetheless, it is important to note that theory-of-mind reasoning can be separated into at least two component processes (see also Tager-Flusberg, 2001): (1) detecting or decoding others’ mental states based on immediately available observable information and (2) reasoning about those mental states in the service of explaining or predicting others’ actions. Examples of detecting or decoding mental states might include some of the relatively rudimentary skills, such as identifying others’ focus of attention based on eye gaze. In contrast, examples of reasoning about mental states within the context of explaining or predicting action may include some more complex aspects of theory of mind reasoning, such as distinguishing jokes from lies, or predicting story characters’ behavior on the basis of false beliefs.

Ordinarily, these two aspects of theory-of-mind work in concert to produce reliable judgments about others’ mental states. However, the distinction between these two processes is important because they each rely primarily on different kinds of social-information processing skills. Mental state decoding relies principally on social information that is gleaned in the immediate and observable environment (e.g., the person’s actions, tone of voice, facial expression, etc.). By contrast, reasoning about others’ mental states to explain or predict action requires one to access knowledge and facts about either the person in question, or their contextual circumstances. For example, in order to correctly infer that someone is sad because she got a poor mark on an exam one needs to detect sadness from the observable information, know that she received a poor mark, and perhaps know that she had wanted to do well.

As was noted above, there is some evidence to suggest that high-functioning individuals with autism have difficulties with both mental state decoding (e.g., poor at decoding mental states relative to gender) and mental state reasoning (e.g., poor at reasoning about beliefs relative to photographs). The fact that the two component processes involved in theory-of-mind reasoning may be affected in autism has interesting implications for understanding the neuro-cognitive bases of theory-of-mind reasoning. One possibility is that detecting and reasoning about mental states are subserved by partially overlapping neural circuitry. The overlapping regions may represent a circumscribed sub-circuit that is specialized for representing information relevant to “theory-of-mind” no matter what the specific task (see Frith & Frith, 1999). If this were the case, this “theory-of-mind” site may be a key to understanding the neuropathology of autism. An alternative possibility is that, as
is the case with language (Neville, Mills, & Lawson, 1992; Pulvermüller, 2000), the different components of theory-of-mind reasoning may be subserved by distinct and non-overlapping circuitry. Individuals with autism may be impaired on both components of theory of mind for one of two reasons: (1) a fundamental deficit in both systems or (2) a core deficit in one component—such as, mental state detection—that affects appropriate development of the other.

Below, we will review evidence from our laboratory in support of the hypothesis that different components of theory-of-mind reasoning are associated with distinct, non-overlapping neural systems. In particular, we will review evidence suggesting that decoding others’ mental states is associated with anterior frontal systems, most likely related to orbitofrontal/medial temporal circuit. Moreover, these systems may be lateralized to the right hemisphere. In contrast, we will review evidence suggesting that reasoning about mental states is related to a relatively circumscribed area of medial prefrontal cortex, that may be lateralized to the left hemisphere. After reviewing this evidence, we will discuss the extent to which our findings regarding the neural circuitry of theory-of-mind relates to current hypotheses regarding the developmental neuropsychological underpinnings of autism (Dawson et al., 2002).

1.2. Neural correlates of decoding mental states

To date, very few studies have been designed specifically with the goal of investigating the neural bases of simply decoding mental states (see Haxby, Hoffman, & Gobbini, 2002, for a recent review). However, some evidence bearing on this issue might come from studies investigating the neural bases of facial emotion recognition. Emotion is a mental state, and judgments about others’ emotions can be made solely on the basis of immediately available information. Neuropsychological evidence suggests that bilateral amygdala damage impairs recognition of emotional states, especially fear (e.g., Adolphs, Tranel, Damasio, & Damasio, 1994; Calder et al., 1996). A role for medial temporal structures such as the amygdala in emotion identification has also found some support in a small-scale MEG study (Streit et al., 1999). In addition to the amygdala, there is some evidence that individuals with ventral-medial prefrontal lesions (including orbitofrontal cortex) are impaired in their abilities to decode emotions from facial expressions (Hornak, Rolls, & Wade, 1996).

These studies have led to the intriguing hypothesis that a circuit including the orbitofrontal cortex and medial temporal regions (including the amygdala) may be crucial for decoding others’ emotional mental states (Baron-Cohen, 1995; Rolls, 1996). However, interpretation of these studies is limited somewhat by their small scale, and by the fact that the stimuli used for emotional judgments were different from ones used for control judgments (e.g., gender, identity). A notable exception comes from Baron-Cohen et al. (1999), who attempted an fMRI investigation of emotional mental state decoding in which gender judgments of the same stimuli were used as a control condition. Unfortunately, these authors noted that there were a number of difficulties with the precise design of their study that limited interpretations of their results.

1.2.1. ERP study: Decoding mental states

We recently completed a study aimed at delineating the spatial and temporal aspects of the neural systems that are important for decoding others’ emotional mental states (Sabbagh, Moulson, & Harkness, in press). To do this, we adapted the “Reading the Mind in the Eyes” task (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) for use in a dense-array ERP paradigm (see Fig. 1). In this modified version, 18 normal participants first saw a word labeling either a mental/emotional state (e.g., embarrassed) or a gender (e.g., female). The word was then replaced with a fixation point, and then a picture of eyes. Participants’ task was to judge (yes/no) whether the eyes depicted the emotion or gender labeled by the preceding word. Event-related potentials were recorded time-locked to the onset of the
picture of the eyes, as participants made their decisions (see Fig. 2). Because emotional mental state and gender judgments were made with respect to the same photographs, differences in ERP activity could not be attributed to general differences between stimuli.

Using this procedure, we found two aspects of the ERP that distinguished mental state decoding from gender decoding. The first component to show the condition effect was the N270-400 that was present over frontal and anterior temporal areas. We found that at right hemisphere sites, the N270-400 component elicited in the mental state decoding condition was more negative than in the gender decoding condition (see Fig. 3). The N270-400 difference was not detected at left hemisphere sites. The second ERP component to show a condition effect was the P300. Specifically, the P300 elicited in the mental/emotional state condition was more positive than in the gender condition. The frontal N270-400 differences and the P300 differences were the only ones to meet significance criteria.

We conducted a source estimation analysis using the low-resolution electromagnetic tomography (LORETA) technique to identify candidate neural generators contributing to the differences in scalp potentials at 320 ms post-stimulus. While this technique is limited by relatively low spatial resolution, previous research has demonstrated that it can be effective in providing highly accurate description of the cortical regions responsible for a given effect (Pascual-Marqui, 1999). This analysis identified two cortical regions that made strong contributions to the scalp electrical differences: orbitofrontal cortex and medial temporal cortex. While each of these areas showed some bilateral activation within each region, higher levels of activation were found in the right hemisphere.

Of course, there are a number of problems with source localization using ERP, and as such, these analyses must be considered cautiously. It is noteworthy, however, that these ERP findings dovetail nicely with the neuropsychological data suggesting that orbitofrontal and medial temporal lesions disrupt emotional recognition. Thus, these findings do provide initial evidence that the neural circuitry underlying one aspect of theory-of-mind reasoning—decoding others’ mental states—may rely on neural circuits within orbitofrontal and medial temporal regions.

1.3. Neural correlates of reasoning about mental states

Although the study described above represents one of the first attempts to understand the neural correlates of
decoding others’ mental states, there is a fairly large body of neuroimaging studies aimed at delineating the neural correlates of reasoning about others’ mental states (see Frith & Frith, 2001, for a review). Typically, these studies have relied on tasks in which the participants’ neural activation is assessed as they perform tasks that require reasoning about mental states as compared with tasks that are closely matched, but do not require reasoning about mental states. For instance, Fletcher and colleagues (Fletcher et al., 1995) developed a PET paradigm in which participants were asked to predict or explain behavior on the basis of either physical causality or story characters’ mental states. Similarly, Goel and colleagues (Goel, Grafman, Sadato, & Hallett, 1995) asked participants to make judgments about whether objects would be considered familiar to either oneself or an historical figure (e.g., Christopher Columbus). Each of these studies, and a number of others that came later (see Siegal & Varley, 2002), have converged on the result that medial frontal areas, particularly in the left hemisphere, make critical contributions to reasoning about others’ mental states.

When taken together with our findings on the neural systems important for decoding others’ mental states, these findings seem to provide some evidence in support of our hypothesis that decoding and reasoning about mental states are subserved by non-overlapping neural systems. Specifically, orbitofrontal and medial temporal areas may be important for decoding mental states, while medial frontal areas are important for reasoning about mental states. Before drawing this conclusion, however, it is important to consider some other interpretations of the mental state reasoning studies. In each of these studies, brain activation was demonstrated as blood flow changes induced as participants reasoned about mental state representations (like beliefs or person-specific knowledge) relative to a control condition in which participants reasoned about physical causality or a real-world state of affairs. Arguably, then, the “mental state” condition places more emphasis on the ability to reason about abstract representations (e.g., beliefs) relative to the other conditions. Thus, it remained possible that left medial frontal cortex may be particularly important for making judgments about abstract representations, relevant to theory-of-mind or otherwise.

This is a particularly important concern because, as discussed above, individuals with autism have special difficulties in reasoning about mental representations (e.g., beliefs) while they are unimpaired in their ability to make judgments about non-mental representations (e.g., photographs).

### 1.3.1. Reasoning about mental states: ERP study

With this in mind, we undertook a study that was designed to investigate whether left prefrontal regions make a greater contribution to reasoning about mental representations relative to non-mental representations (see Sabbagh & Taylor, 2000, for fuller description of this work). Our study took as its starting place the false belief/false photograph dissociation apparent in autism. In our study, 23 typically developing adults had to reason about outdated (or false) beliefs and photographs. These beliefs and photographs were outdated in the sense that the scene changed subsequent to the creation of the representation, and as such, the representation no longer matched the scene. Participants read a belief and a photo version of 40 stories for a total of 80 stories (see Table 1 for an example). Following each story, participants were asked a series of control questions to ensure that they understood the story, and a test question asking them to judge the contents of the representation. The test question was constructed such that participants could not answer until the final word appeared. ERPs were recorded to the onset of the final word in the question. The events associated with the test question and recording of ERPs are illustrated in Fig. 4.

The key results of this study are illustrated in Fig. 5. Even in this tightly controlled comparison, reasoning about mental states was uniquely associated with left frontal activity beginning 300 ms following the onset of the stimulus. Specifically, the ERP was characterized by a slow-wave effect, which was more positive in the belief condition relative to the photograph condition. This slow-wave condition effect was remarkably focal—it was present only on a cluster of four sites over lateral frontal areas. This effect was mirrored at left parietal sites where a slow-wave effect also existed, but in the opposite direction: the ERP elicited in the belief condition was more negative relative to the photograph condition. This pattern of differences was maintained throughout the recording epoch, and was most distinct at around 820 ms.

Unfortunately, we were unable to conduct source localization analyses on these data, because precise electrode locations were not measured. Nonetheless, the

### Table 1

<table>
<thead>
<tr>
<th>Example of story from mental state reasoning task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Belief</strong></td>
</tr>
<tr>
<td>Fred put a flower and a bottle of wine on the table</td>
</tr>
<tr>
<td>Fred showed them to his friend Marianne</td>
</tr>
<tr>
<td>Marianne thought they looked pretty</td>
</tr>
<tr>
<td>Then, Marianne fell asleep in her room</td>
</tr>
<tr>
<td>While Marianne was asleep, Fred moved the flower</td>
</tr>
<tr>
<td>Fred put the flower on the piano</td>
</tr>
<tr>
<td>He left the bottle of wine on the table</td>
</tr>
</tbody>
</table>
pattern of differences is consistent with a left anterior medial frontal generator that is suggested by the fMRI and PET studies reviewed briefly above. These findings find converging evidence for the importance of medial prefrontal regions of the left hemisphere in reasoning about others’ mental states.

2. Summary and discussion of ERP findings

One of the most striking aspects of our ERP studies is that, when viewed together, there was essentially no reliable overlap between the neural systems that were associated with decoding versus reasoning about mental states. In particular, the ERP differences that indexed mental state decoding were localized to inferior frontal and anterior temporal regions of the right hemisphere. Source localization analyses estimated that these ERP differences may have been generated within orbitofrontal prefrontal and medial temporal regions of the right hemisphere. In contrast, reasoning about others’ mental states was associated with ERP activity that was focal over prefrontal areas of the left hemisphere. Although we did not conduct source localization analyses on these data, the pattern of activity was consistent with a medial frontal generator within the left hemisphere that has been described in fMRI and PET studies of mental state reasoning (see Frith & Frith, 2001).
These findings support the notion that no single brain area is active when humans engage theory-of-mind relevant informational content. Instead, the neural systems that support successful theory-of-mind judgments are distributed across different cortical regions. These findings raise important questions regarding how to characterize the contributions of these different neural circuits. One possibility is that instead of there being one specialized circuit for reasoning about mental states, there may be several—each one dedicated to carrying out a specific step in the process of making appropriate theory-of-mind judgments. For instance, there may be a specialized pathway within the orbitofrontal/temporal circuit that is specifically dedicated to decoding mental states. It seems possible that one aspect of this specialization may be its laterality. There are a number of findings now that suggest that decoding others’ mental states from observable information may be a skill that is lateralized to the right hemisphere (Sabbagh, 1999; Sabbagh et al., in press). Complementing this right lateralized orbitofrontal/temporal circuit, there may be a specialized region of medial frontal regions of the left hemisphere which may be especially important for reasoning about mental states. This hypothesis gains support from the fair amount of consistency across studies that have used tasks that require theory-of-mind reasoning and revealed activation of left medial frontal areas (Frith & Frith, 1999; Siegal & Varley, 2002).

A second possibility is that instead of being specialized, each of these circuits performs a domain-general computation that is required when engaging different theory-of-mind related processes, but not specific to this social-cognitive domain. For instance, orbitofrontal areas may be important not only for decoding mental states but also for decoding the meaning of stimuli in other non-social domains. Similarly, the left medial frontal contributions could reflect a domain-general cognitive contribution that happens to be heavily taxed when participants are engaged in theory-of-mind reasoning.

Our studies do not provide us with the opportunity to determine whether we have indexed domain-specific or domain-general neural systems. However, some insight on this issue may be taken from examining the tasks that we used—and that have been used with individuals with autism—to reveal electrophysiological dissociations associated with theory-of-mind. In both the case of mental state decoding and mental state reasoning, the tasks were chosen because the mental state and non-mental state conditions are, at least on their surfaces, very well matched in terms of peripheral domain-general cognitive demands. For example, the cognitive demands in the false belief task seem very similar to those required in the false photograph task (e.g., working memory, logical reasoning, etc.). Similarly, there is no a priori reason to suspect that the information-processing demands associated with decoding mental states from pictures of eyes would be quantitatively and qualitatively different from those required to decode gender. From this basis, then, it seems reasonable to suggest that our electrophysiological differences reveal the operation of neural circuits functionally dedicated to decoding and reasoning about mental states (e.g., Frith & Frith, 2001).

It is important to note, however, that the assumptions regarding how well these experimental tasks are matched have not been assessed directly. Thus, it remains possible that the neural systems we suggest are crucial for decoding and reasoning about mental states may be related to domain-general cognitive skills that are more heavily tapped in course of these processes. Although no research has investigated this possibility directly, it is possible that the findings from our ERP studies may provide some hypotheses about candidate cognitive skills that may be especially important for theory-of-mind reasoning. For instance, reasoning about mental states may be rely on contributions from left prefrontal areas because this skill may be related to executive functioning and “inhibitory control.” A number of researchers have argued that prefrontal regions are associated with executive functions including working memory (Smith & Jonides, 1998), error monitoring (Gehring & Knight, 2000), and inhibitory control (Diamond, 1998). Moreover, recent research strongly suggests that these executive skills are crucial for reasoning about others’ mental states (Carlson & Moses, 2001; Russell, 1996). Together, these findings strengthen the possibility that left frontal contributions to theory-of-mind may be related to executive functions in children. One caveat is that the bulk of research from both animal and human work suggests that the aspects of executive function that are typically related to theory-of-mind performance in children are located with in dorsal–lateral prefrontal cortex as opposed to medial frontal cortex. Clearly, more research is needed to better understand this discrepancy.

In a similar vein, decoding mental states may recruit neural circuitry within orbitofrontal contributions because these areas seem to play a special role in decoding the content of emotionally reinforcing stimuli generally (Rolls, 1996, this issue). A considerable amount of research now suggests that the orbitofrontal cortex plays a crucial role in decoding the content of gustatory and olfactory stimuli (see Rolls, 1996, for a review). Similarly, Bechara and colleagues have demonstrated that individuals with damage to orbitofrontal cortex have difficulty with so-called gambling tasks—card games in which they are required to distinguish “rewarding” decks from “punishing” decks (e.g., Bechara, Tranel, & Damasio, 2000). Arguably, the emotional value of social interaction crucially relies upon the ability to decode others’ mental states. Thus, when mental states are de-
coded, they may be represented primarily as social emotional reinforcers. It may be, then, that orbitofrontal contributions to mental state decoding may be related to the emotionally reinforcing character of mental states themselves.

As was noted above, we cannot marshal strong evidence in favor of either a domain-specific or domain-general interpretation of our results regarding the neural bases of decoding or reasoning about mental states. One thing that is clear, however, is that neural systems associated with either component of theory-of-mind are distinguishable from those that are associated with similar judgments in other domains.

3. Neurobiology of autism: Is there a core deficit?

From the outset, a primary motivation for conducting our studies was to better understand the neuro-pathological factors that contribute to autism. This is an important aim because although it is generally recognized that the cognitive and social deficits seen in autism have their roots in neurodevelopmental abnormalities, researchers have had considerable difficulty in accurately characterizing specific sites of pathology (Minshew & Goldstein, 1998; Piven & O’Leary, 1997). More specifically, neuroimaging studies have failed to converge on a single focal defect that characterizes the autistic brain (see Rumsey & Ernst, 2000, for a recent review). There are likely a number of reasons for this lack of convergence. For one, autism is a heterogeneous disorder with respect to presentation characteristics. Affected individuals can range from having profound IQ deficits to relatively normal intelligence. Or, affected individuals may present either with or without epileptiform symptoms. These relatively large variations may obscure more subtle common neural characteristics among individuals.

A second, and perhaps more important, reason for the lack of convergence is that autism is a developmental disorder. Building on recent work in the domain of developmental cognitive neuroscience, Karmiloff-Smith (1997) argues that development itself plays an important role in shaping the functional organization of neural systems. As such, it does not necessarily follow that individuals affected by a given developmental disorder would have a characteristic brain lesion. Instead, the whole pattern of functional brain organization must be considered in light of abnormal initial states (e.g., Deb & Thompson, 1998; Elman et al., 1996). This shift in perspective militates for understanding the neuro-cognitive bases of autism from a fundamentally developmental perspective.

The theory-of-mind framework may provide an excellent route to adopting just this perspective. Delineating the developmental trajectory of theory-of-mind skills from infancy through the preschool period has been a central concern for many researchers in the area of cognitive development. Although even a partial review of this literature is beyond the scope of this paper, one consistent finding is that children are able to detect or decode others’ mental states before they are able to reason about those mental states in a sophisticated manner. One example of this progression is the case of eye-gaze monitoring. Infants as young as 6-month-old show evidence of attending to others’ direction of eye-gaze (Hood, Willen, & Driver, 1998), although it is not until the age of 18-months (Baldwin, 1991) or later (Lee, Eskritt, Symons, & Muir, 1998) that they are able to reliably use this information to reason about others’ intentions in a sophisticated manner. Research suggests that this progression may also hold for later developing aspects of theory-of-mind reasoning, such as understanding of knowledge—even young children seem to be able to reliably indicate whether someone is knowledgeable (Bartsch & Wellman, 1994; Pratt & Bryant, 1990), but they have difficulty reasoning on the basis of that information (Wimmer, Hogrefe, & Perner, 1988).

That decoding emerges prior to reasoning in the theory-of-mind domain has important implications for autism. Specifically, given that autism is a developmental disorder, it may be the case that at the core of this disorder is the ability to engage the right orbitofrontal/medial temporal circuit that we hypothesize makes an important contribution to mental state decoding. This core deficit may, in turn, lead to additional abnormalities in functional brain organization, and theory of mind reasoning. A key piece of evidence in evaluating this question comes from determining whether young children with autism are specifically impaired on tasks designed to tap orbitofrontal/medial temporal functioning relative to other anterior systems.

Unfortunately, the evidence in this regard is sparse. Dawson and colleagues (Dawson et al., 2002) recently reported that autistic and non-autistic children who are matched for mental age do not differ from one another on tasks that are typically thought to tap dorsolateral versus orbitofrontal functioning (Dawson et al., 2002). However, this same group of researchers has also reported two very intriguing findings that indirectly link variability in orbitofrontal functioning to both autism and to theory-of-mind. First, Dawson and colleagues (Dawson, Meltzoff, Osterling, & Rinaldi, 1998; Dawson, Osterling, Rinaldi, Carver, & McPartland, 2001) have shown that individuals with autism are relatively more impaired than mental age matched controls on the delayed non-matching to sample task, which taps the ability to form rules regarding the links between stimuli and rewards. This is important because research has shown that this skill is mediated through ventromedial prefrontal cortex, including orbitofrontal cortex (Meunier, Bachevalier, & Mishkin, 1997). Second, across all
participants, tasks associated with orbitofrontal functioning was more likely to be impaired in the ability to establish “joint attention,” which requires the ability to detect others’ attentional and intentional states based on the direction of their eye gaze. Developmentally, children’s abilities to establish joint attention is widely regarded as one of the first milestones on the road to acquiring a theory-of-mind (Moore, 1999). Moreover, it is gaining validity as one of the earliest warning signs of autism (Baird et al., 2000). These findings linking joint attention abilities with ventral-medial prefrontal function provide converging evidence for both (1) the fundamental role of ventral-medial regions in decoding mental states (including joint attention), and (2) the possibility that aspects of orbitofrontal dysfunction may play a crucial role in understanding the developmental neuropsychological factors that lead to the social deficits evidenced by individuals with autism.

Finally, it should be noted that this hypothesis regarding the central importance of the orbitofrontal/medial temporal circuit for both social cognition and autism extends a long standing hypothesis regarding the pathological role of the hippocampus and amygdala in autism. Bachevalier (1994) and others have argued from the basis of neuropsychological, animal lesion, and neuroimaging studies that the medial temporal lobe is an important source of dysfunction in autism (see also Howard et al., 2000). By including a role for orbitofrontal regions, we may gain important insight into the neuro-cognitive bases of the theory-of-mind deficits associated with autism.

4. Conclusions and potential clinical implications

In summary, our findings suggest first that there is no one highly circumscribed neural circuit that is active any time one engages in theory-of-mind relevant reasoning. Instead, the neural circuitry that underlies the ability to decode mental states from facial displays differs from that underlying the ability to reason about mental states. Specifically, orbitofrontal cortex may be especially important for decoding mental states, while left medial frontal regions may be important for reasoning about mental states. It is presently unclear as to whether a domain-specific or domain-general interpretation of these neuro-cognitive contributions is most warranted. Nonetheless, these findings may provide a springboard for exploring some hypotheses about the neuropathological bases of autism. Specifically, it seems possible that autism may be characterized by a core deficit in the functioning of the orbitofrontal/medial temporal circuit that may underlie their difficulties in decoding mental states. Given that decoding mental states is widely considered to be the earliest emerging building block of theory-of-mind development, it could be that a fundamental abnormality in mental state decoding is the basis of social cognitive deficits in autism.

We would like to conclude with a speculation about the possible clinical implications of our findings. A principal aim of therapists working with individuals with autism is to improve their social functioning skills. We, along with others, have hypothesized that the “core deficits” associated with autism lie in the orbitofrontal/medial temporal circuit that is important for decoding mental states. In contrast, the left prefrontal regions that are important for reasoning about mental states may be relatively less affected. If this were the case, it is possible that individuals with autism may reveal strengths in their ability to reason about mental states if the links between mental states and their observable correlates (e.g., facial expression) were made more transparent. Put in another way, perhaps an effective therapeutic strategy would be to help individuals with autism to recognize mental states by presenting their contents more transparently. For instance, individuals with autism might have difficulty recognizing that someone is sad, but if told explicitly, may be able to make appropriate judgments regarding that person’s attendant mental states.

There is some evidence to suggest that this might be the case. For instance, Wellman et al. (2002) have shown that when others’ mental states are made explicit through the use of cartoon-like “thought bubbles” individuals with autism can solve basic theory-of-mind tasks, including false belief. Similarly, Tager-Flusberg (2001) reviews evidence suggesting that individuals with autism can use language-based strategies to “hammer out” solutions to problems that require reasoning about others’ mental states. More research is required to determine whether these strategies lead to lasting improvements in social cognitive skills in individuals with autism. Nonetheless, this approach may ultimately be an effective way of capitalizing on the neuro-cognitive strengths of individuals with autism to better their everyday social cognitive skills.

Acknowledgments

This article was prepared with support of an NSERC Discovery Grant. The author thanks Sid Segalowitz and Phil Zelazo for very helpful comments on a previous draft.

References


Pratt, C., & Bryant, P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development, 61*, 973–982.


