BRAIN RESEARCH XX (2006) XXX–XXX

available at www.sciencedirect.com

SCIENCE DIRECT®

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Why and how to study Theory of Mind with fMRI

*Rebecca Saxe**

Harvard Society of Fellows, Harvard University, Cambridge, MA 02138, USA
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Accepted 1 January 2006

Keywords:
Theory of Mind
fMRI
Development
Narrative
Regions of interest

ABSTRACT

Social cognitive neuroscience investigates the psychological and neural basis of perception and reasoning about other people, especially in terms of invisible internal states. This enterprise poses many challenges. The current review describes responses to three such challenges: deriving hypotheses from developmental psychology, using verbal narratives as stimuli, and analysing the results in functionally defined regions of interest.

1. Introduction

> "You don't look sad now, he thought. And he wondered what she was reading, and exaggerated her ignorance, her simplicity, for he liked to think that she was not clever, not book-learned at all. He wondered if she understood what she was reading. Probably not, he thought."~Virginia Woolf, To the Lighthouse.

Fiction depends on an ordinary miracle: given just a few sentences, we perceive whole human minds at work, their flow of thoughts and feelings, perceptions and self-deceptions. In this respect, reading fiction is representative of the most social cognition. We constantly "go beyond the data", making inductive leaps from sketchy evidence to rich interpretations of the people around us. Such inferences are the bedrock of our daily lives. Even 1-year-old infants, when they are learning their first words, rely on sophisticated inferences about the speaker's intentions (Baldwin, 1993; Bloom, 2000). Losing such social fluency is more devastating to individuals and their caretakers than the loss of memory, hearing, sight or speech (Levenson, R., personal communication).

In spite of the importance of reasoning about other minds in everyday life, neuroscientific research in this area is in its infancy. Do we have special mechanisms, designed by evolution for recognising and/or reasoning about other minds, or does social cognition share the general-purpose machinery we use for recognising chairs and reasoning about falling apples? How and why does the human brain succeed so easily where computers and logicians fail? What neural structures underlie our inductions about other people's internal states? The present special issue recognises the importance of these open questions and surveys opinions about and solutions to the challenges that these questions pose.

At least three fundamental challenges face neuroscientists interested in person perception (or social/cognitive psychologists interested in neuroscience): (Q1) which hypotheses to test, (Q2) which stimuli to use and (Q3) how to analyse the results. Of course, the solutions must be case-specific, depending on the goals of each individual study. In what

* *Department of Brain and Cognitive Sciences, MIT 46-4019, Cambridge, MA 02138, USA.*
E-mail address: Saxe@mit.edu.

0006-8993/$ – see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.brainres.2006.01.001

follows, I will advocate three particular approaches that have enriched my own research: (A1) deriving hypotheses from developmental psychology, (A2) using verbal narratives as stimuli, and (A3) analysing the results in functionally defined regions of interest (fROIs). As such, this review considers the questions behind current research on Theory of Mind, and not the questions answered by it; for a review of recent results, see Saxe (in press).

2. Theoretical frameworks: developmental psychology

The neuroscience of person perception is in many respects a brand new field, susceptible to the usual pitfalls of immaturity. As this special issue attests, "social cognitive neuroscientists" do not yet agree on a central set of phenomena that theories of person perception should explain or on the core processes to which such explanations should appeal. Thus, the first challenge for social cognitive neuroscience is to formulate the hypotheses that we can then set out to test. One promising approach is to begin by borrowing ideas from neighbouring disciplines, including social psychology, neuropsychology, animal behaviour, ethics and even economics. My own first steps in social cognitive neuroscience have been most informed by the theoretical and methodological traditions of developmental psychology, in particular, under the rubric of the development of "Theory of Mind" (e.g. Wimmer and Perner, 1983; Flavell, 1988; Perner, 1993).

The most intensely studied phase in the development of Theory of Mind occurs between ages three and five. The contour of this transition is well illustrated in an elegant series of experiments by Ziv and Frye (2003). Short stories about animal characters ('Duck' and 'Cat') and their belongings were enacted for 3, 4 and 5-year-olds. In the change-of-locations task, Duck put his ball in a bag. While Duck was away (where he could not see or hear what Cat was doing), Cat took the ball out of the bag and put it into an envelope. Then, the child was asked two kinds of question in counterbalanced order: belief questions (e.g. "Where does Duck think the ball is?") and desire questions (e.g. "Where does Duck want the ball to be?"). In the change-of-contents task, a similar scenario unfolded, except that Cat took out the ball and put a book into the bag, so that both the relevant beliefs and desires concerned the contents of the bag.

For both change-of-location and change-of-contents tasks, 3-year-olds performed just like 5-year-olds on desire questions: over 80% of children at all ages concluded that Duck wanted the ball to be in the bag (as he left it). Three-year-olds succeeded on versions of each task in which the hidden object was neutral, desirable or undesirable. Furthermore, children of all ages were able to infer the character's desire from his behaviour, even when pitted against the child's own preference (i.e. if Duck put a ball in the bag, then he wants the ball to be in the bag, even if what Cat has placed in the bag is a more 'desirable' chocolate).

The striking developmental change appeared on the belief questions. Five-year-olds judged that Duck's belief still reflected the original situation: the scene as it was when Duck left the room. That is, 5-year-olds correctly attributed a false belief. By contrast, 3-year-olds scored reliably below chance on belief questions; the younger children claimed that Duck's belief reflected the actual (real) situation rather than the original situation in both change-of-location and change-of-contents tasks (see also Wellman et al., 2001).

Children's changing understanding of belief is evident on a range of related tasks. The crucial difficulty seems to lie in understanding how a (truly held, full conviction) belief or perception can nevertheless represent a state of affairs that is different from the real one. Children who fail "False Belief" tasks—like the change-of-location and change-of-contents tasks used by Ziv and Frye (2003)—also fail to distinguish between what an object looks like and what it really is (Gopnik and Astington, 1988). When 3-year-olds are shown that an object that looks like a rock is really a sponge, they later answer both that it "really is" a sponge and that it also "looks like" a sponge. Relatedly, these children have difficulty both deploying, and understanding, deception (e.g. Wimmer and Perner, 1983). When a goblin (in a fairy tale) dresses up as a boy in order to deceive a girl into playing with him, 3-year-olds say that the goblin will not play mean tricks on the girl because "he is dressed up as a boy and he is nice now" (Peskin, 1993, described in Perner, 1993). The majority of 3-year-olds also assert that the girl thinks the goblin is pretending to be a boy, and not that he really is a boy.

What, then, is the conceptual advance that distinguishes 5- from 3-year-olds? It would certainly be misleading to claim that 5-year-olds "have" a Theory of Mind, whereas 3-year-olds do not have one (Bloom and German, 2000). The 3-year-olds studied by Ziv and Frye (2003) were as good as 5-year-olds at inferring a character's desires from his behaviour and gave systematic (albeit wrong) answers to the belief questions, suggesting that these children do not hesitate to explain and predict people's actions based on inferred mental states. Moreover, even much younger children make inferences about people's intentions, perceptions and emotions and the interrelations among these mental states (e.g. Harris et al., 1989; Repacholi and Gopnik, 1997; Phillips et al., 2002). For example, 2-year-old children understand the basic relationship between desires and emotions. Given a story about a boy who wanted a puppy and then got one, 2-year-olds choose a happy face to show how the boy feels, but they choose a sad face if the boy who got a puppy had wanted a bunny (Wellman and Woolley, 1990; Wellman et al., 1995).

Rather, 3-year-olds lack one specific component of a 5-year-old's Theory of Mind: a clear distinction between *what* a person's mental state is about (the state of affairs to which the belief or perception refers) and *how* that state of the affairs is represented (what the person believes or perceives to be true of it; Perner, 1993). This distinction allows the older children to understand how people's mental representations of the world may differ from the way the world really is. As a result, 5-year-olds are sometimes said to have (and 3-year-olds to lack) a *representational* Theory of Mind.

Still, many questions about the trajectory of Theory of Mind development remain unanswered; these open problems provide an opportunity for social cognitive neuroscientists. The direct approach to these questions would benefit from scanning the brains of children, but, in fact, many hypotheses derived from developmental psychology make distinct

predictions about neural divisions of labour even in the adult brain. For example, one such question is whether early- and late-developing components of Theory of Mind rely on common or distinct psychological or anatomical substrates (see also Saxe et al., 2004a). Reasoning about beliefs develops later than an earlier Theory of Mind that includes attribution of desires, perceptions and emotion. Does the later emerging competence colonise the same neural systems that underpin earlier reasoning? If so, we would predict that attributions of early-developing concepts, for instance, would recruit the same brain regions as belief attribution. If, on the other hand, reasoning about beliefs draws on distinct systems or abilities, then reasoning about concepts that 3-year-olds have mastered should not produce activity in regions associated with belief attribution and might recruit a distinct set of brain regions even in adults (Saxe and Powell, in press).

A related question concerns the whether changing performance on False Belief tasks reflects a developmental change specifically in the domain of Theory of Mind (e.g. acquiring a previously absent representational concept of belief; e.g. Wellman et al., 2001) or instead reflects improvements in some other capacity that is necessary for good performance on the task, such as increasing inhibitory control or other "executive" skills (e.g. Moses, 2001; Carlson and Moses, 2001). To answer a belief question in the change-of-location task, a child must be able to juggle two competing representations of reality (the actual state of affairs and the situation represented in the protagonist's head) and to inhibit an incorrect but compelling answer (the actual location of the object). Young children's performance is similarly delayed by inhibitory demands on tasks that do not tap Theory of Mind at all, such as reasoning about non-mental false representations (e.g. false photographs or maps, Zaitchik, 1990).

Two alternative interpretations could explain the relationship between inhibitory control and reasoning about beliefs, each making distinct predictions for social cognitive neuroscience. One possibility is that inhibitory control plays a constitutive role in reasoning about beliefs (relative to desires, perceptions and emotions) and that the developmental transition in Theory of Mind performance reflects not a domain-specific mechanism for understanding belief, but just the increase of inhibitory control with age. On this hypothesis, there would be no domain-specific brain regions for the *representational* Theory of Mind, but only regions for social cognition (relatively constant over development) and for inhibitory control generally (changing with age). Alternatively, inhibitory control may just facilitate children's early learning about beliefs and the construction of a domain-specific representational Theory of Mind. In this case, inhibitory control (and associated brain regions) would not be intrinsic to belief attribution. Evidence for a domain-specific brain region, recruited during reasoning about mental representations but not during reasoning about non-mental representations, or about non-representational mental states, would support this latter hypothesis: that the late-developing representational Theory of Mind has a domain-specific neural (and psychological) foundation (Saxe and Kanwisher, 2003). Similar arguments concerning the role of language in the development of a representational Theory of Mind will be discussed in the next section.

The same neuroscientific investigations may also contribute to the theoretical debate within developmental psychology over whether the developmental trajectory of Theory of Mind is best described as the maturation of one or more Theory of Mind "modules", or as the elaboration of a quasi-scientific theory, or some combination of the two. Theories and modules differ most clearly in the proposed mechanism (and process) of conceptual change. Modules are usually envisioned as developing "along an internally directed course under the triggering and partially shaping effect of the environment" (Chomsky, 1980). Informational limitations and the genetic endowment ensure that external influence is relatively narrow and constrained (Fodor, 1983), keeping the modular structure consistent across individuals. Theories are more deeply susceptible to the influence of evidence and the environment. Consequently, the fate of an outgrown concept may be different in a module than in a theory. Modules (especially for 'core knowledge') are conceived as enduring over development (Carey and Spelke, 1994). Developmental change occurs by modifying the interpretation of the module's output (Scholl and Leslie, 1999), not by destruction or modification of the module itself. Theories, on the other hand, can be altered directly. Unsatisfactory concepts or theories are replaced or changed. Thus, evidence that the neural substrate of the early-developing components of Theory of Mind remain unchanged—and, in the adult brain, distinct from the neural correlates of late-developing components—may provide evidence that the earlier, outgrown mechanism is not replaced or colonised, consistent with modularity.

Third, within the early-developing component of Theory of Mind, the question remains open to what extent understanding desires, perceptions and emotions rely on distinct or common psychological or anatomical substrates. Baron-Cohen (1994) has divided this domain into two distinct components. The 'Intentionality Detector' represents behaviour in terms of goals, while the 'Eye Direction Detector' detects eyes and represents the direction as the Agent 'seeing'. By contrast, Leslie (1994) groups understanding of action, goals and perceptions together within a single module. Consistent with Leslie's proposal, in a longitudinal study of 9- to 15-month-olds, Carpenter et al. (1998) found that the emergence of gaze following (attribution of perception) and imitation of novel actions (goal attribution) were positively correlated with each other and uncorrelated with concurrent non-social developments such as object permanence. The same hypotheses could be tested in the adult brain. Near the posterior superior temporal sulcus, different studies have reported brain regions recruited during observation of gaze shifts (Puce et al., 1998, Pelphrey et al., 2003) and of intentional actions (Decety and Grezes, 1999), but no study has yet directly compared these two conditions. Neuroimaging work along these lines may help to determine whether attributions of perceptions and desires/goals rely on the same brain regions, consistent with Carpenter et al.'s (1998) conclusion that these two parts of early Theory of Mind reflect "the same underlying phenomenon."

Finally, the controversy within developmental psychology that has received the most attention from social cognitive neuroscientists recently is the debate between Theory-theorists (usually understood broadly to include modules) and

Simulation-theorists over the format of knowledge of other minds (Carruthers and Smith, 1996). The Theory-theory position is, approximately, that children and adults have concepts of kinds of mental states and beliefs about the relationships among those mental states and between mental states and actions. We "go beyond the data" of perceived behaviour by using this theory to actively reconstruct the contents of other minds (and even of the observer's own mind in the relatively distant past or future). These same concepts and beliefs about the mind then support predictions, explanations and verbal descriptions of behaviour and internal experiences (Gopnik and Wellman, 1992). Notably, though, conceiving of a belief may depend on entirely distinct machinery from having that belief. In particular, young children have representational beliefs many years before they have the concept of belief that is necessary for attributing such a belief to themselves or anyone else.

Simulation Theory proposes that people need not use a naive theory of psychology, or indeed any mental state concepts, when predicting and explaining actions. Instead, the observer uses her own mind as a model for another mind (Harris, 1992; Nichols et al., 1996). Simulation Theory has recently been embraced with enthusiasm by neuroscientists and cognitive neuroscientists following the discovery of the "mirror system": neurones (or in humans, brain regions) that are recruited both when performing and when watching someone else perform a particular action (Gallese and Goldman, 1998). Similar systems have also been discovered for some basic emotions (e.g. fear and disgust) and for sensations (Gallese et al., 2004).

Still, the relationship between Simulation and Theory remains wide open for investigation by social cognitive neuroscientists (Saxe, 2005). What is the scope of the mirror system? In what contexts, and for which mental states is Theory-theory a better model? How are theories implemented in the brain? Each of these questions may benefit from neuroscientific investigation.

In all, the 25 years of developmental psychology of Theory of Mind provides effective scaffolding for initial investigations in social cognitive neuroscience. Many of the well-elaborated theories and influential controversies can be tested in neuroimaging studies. At the same time, decades of careful experimentation have lead to an ample collection of empirical paradigms. The next challenge is how to adapt these paradigms and theories for use with adults in fMRI.

3. Sources of social information: verbal stimuli

The fMRI scanner environment is inhospitable to natural social interaction. Not only is the subject immobilised, with his or her head deep inside a thick and very noisy tube, but current analysis techniques almost exclusively require carefully timed, controlled and replicable events. Inevitably, then, fMRI experiments on person perception present subjects with abstracted and impoverished sources of social information. This is the second major challenge for social cognitive neuroscientists: how to evoke social cognitive behaviours in an environment so inconducive to social interaction.

The choice of which source of information to provide—that is, what kind of stimuli to present—reflects both intuitions about the structure of the natural social world and the results of previous experimental paradigms in related fields of psychology. Both factors frequently lead to a focus on face-to-face dyadic interactions. On this view, the paradigmatic exchange of social information occurs between two people looking at each other. The result has been social cognitive neuroscience research on the perception of primarily non-verbal social information: perception and recognition of face identity, facial expression, eye gaze, body posture, body motions (especially reaching actions), race, gender, attractiveness and personality traits (see e.g. Allison et al., 2000).

Undeniably, partners in a dyadic interaction do acquire social information about one another through non-verbal sources. But, humans also rely substantially on another source of social information: verbal communication. The biggest challenge facing anyone who seeks to understand another mind is that mental states cannot be observed in the environment; beliefs and desires, doubts and convictions, moods and motives are all invisible, abstract entities. One invaluable way to learn about these elusive components of the mind is therefore to listen to how other people *talk* about the mind.

Research in developmental psychology, in particular, suggests the importance of language for Theory of Mind (Astington and Baird, 2005). As described in Section 2, the classic test of children's capacity for reasoning about other minds is the False Belief task. Dunn and colleagues (e.g. Dunn and Brophy, 2005) first reported that language ability predicts success on the False Belief task, independent of age. A similar correlation is observed in samples of both healthy children and children with autism and other developmental disabilities (Astington and Jenkins, 1999, Peterson and Siegal, 1999). In a striking example, deaf children of hearing parents (that is, whose parents are non-native signers) are selectively delayed in passing the False Belief task (e.g. Peterson and Siegal, 1999). These children have similar difficulty even on non-verbal tests of False Belief understanding, suggesting that the delay does not reflect the language demands of the tasks themselves (e.g. Figueras-Costa and Harris, 2001). Deaf children of deaf parents (native signers), by contrast, are not delayed (De Villiers, 2005b). Clearly, linguistic exposure influences Theory of Mind development.

The controversial question is: *how* does linguistic exposure influence Theory of Mind development? Some hypotheses focus on the role of syntax (De Villiers, 2000, 2005a). Proficiency with particular grammatical structures (especially sentence complements, such as 'He knows that *the cup is on the table*' or 'She said that *the chocolate was in the box*'), is necessary for forming sentences about some mental states and therefore might be necessary for forming *thoughts* about other minds. This grammatical format is particularly tightly tied to the transition in Theory of Mind development between ages 3 and 5. In English, later-developing mental state concepts, like 'believe' and 'know', require sentence complements, while earlier-developing mental state concepts, like 'want' and 'see', can be attributed without sentence complements (as in 'He sees *the cup on the table*' or 'She wants *the chocolate in the box*'). The conceptual

transition might therefore be entirely parasitic on linguistic development.

Recent evidence, though, suggests that language plays predominantly a communicative rather than a constitutive role in Theory of Mind development. Importantly, the correlation between linguistic exposure and Theory of Mind does not depend on the use of specific grammatical structures, such as sentence complements. In training studies, performance on a False Belief task is enhanced by simply discussing different perspectives on the same event/object, without any use of sentence complements (Lohman and Tomasello, 2003; Harris, 2005). Nor does the trajectory of Theory of Mind development simply follow linguistic boundaries. Across languages, complement structure may be necessary for statements about beliefs but not about desires (as in English), for beliefs and desires (as in German), or for neither belief nor desires (Chinese). Nevertheless, children learning each of these three languages all understand and talk about desires significantly earlier than beliefs (Tardif and Wellman, 2000; Perner et al., 2005). Finally, adults with severe impairments of grammar are not impaired on Theory of Mind tasks (Varley and Seigal, 2000; Varley et al., 2001).

Thus, the evidence suggests that linguistic exposure is critical for Theory of Mind development but not because mental state concepts are *represented* verbally; rather, because verbal communication may be the critical source of evidence children use to *learn* those concepts. The special features of representational mental states, like thoughts and beliefs, may be particularly salient in two verbal contexts: when talking about past events or about third-person narratives involving an absent protagonist (Lagattuta and Wellman, 2001). Both contexts serve to highlight the potential for differences between the current world state and the contents of representational mental states (Harris, 2005). The quantity and quality of such conversation about the mind (and the past) in a young child's environment predict later success on formal tests of understanding other minds (Dunn and Brophy, 2005).

Potentially, fMRI evidence could also contribute to this debate by asking whether the representation of mental state concepts is particularly dependent on "language" regions of the brain. Preliminary evidence suggests that it is not (Saxe et al., 2004a), but much more research is needed.

Notably, social learning does not end with childhood; adults spend a substantial proportion of natural conversations talking about people. Information about the causes and consequences of human actions, and the structure of society, is continually exchanged in informal contexts, like gossip. Diary studies suggest that 80 to 90% of topics in naturally occurring conversation concern the actions (~35%), intentions (~20%) and attitudes (~20%) of people known to the conversants (Emler, 1994). The vast majority of these conversations refer to people or events not present or concurrent; only about 7–8% of conversation topics refer to current states of the speaker's mind and/or body. Eavesdropping studies provide convergent evidence: over 60% of overheard utterances refer to "social topics" (Dunbar, 2004). The tendency to talk about one another also appears to be robustly cross-cultural (e.g. Haviland, 1977).

Verbal communication is thus a ubiquitous source of information about other people, but do adults use this information to continue developing their Theory of Mind? Narratives—stories—are potent sources of social information, easy to learn and then retrieve, recreate and pass on (Johnson, 1993). Moral reasoning and learning, in particular, may rely on stories and story-telling (Goldman, 1993). Rather than being constituted dominantly of general principles or abstract rules, Owen Flanagan (1998) argued that moral knowledge operates through mechanisms for complex pattern recognition, relating current possible or actual actions to learned prototype narratives (see also Bettelheim, 1975, Humphrey, 1997). The prototypes themselves, Flanagan hypothesised, are trained through exposure, including exposure to conversation in society.

Similarly, outside the moral domain, adults acquire from gossip not just a list of transient facts about a specific individual or incident, but rather general information about the structure of the social world. When college students are asked directly what, if anything, they learned from a (self-chosen) interesting anecdote about another person, 93% responded with generalisations about human behaviours that fit into a broader Theory of Mind: "sometimes people live up to self-fulfilling prophecies," "cheerful people are not necessarily happy people," "guys compare their partners," and so on (Baumeister et al., 2004, see also Baxter et al., 2001).

In summary, verbal communications are a naturally occurring and rich source of information about the structure and cause of human thoughts and actions. Verbal information is instrumental in Theory of Mind development; talk about absent third persons and past events highlights differences in perspectives and so illustrates the structure of representational mental states. Furthermore, talk about the mind continues throughout life in both formal and informal contexts. All of these considerations suggest that a verbal narrative about an absent protagonist is a common and natural context for the acquisition of social information. Given this review of the evidence from both developmental and social psychology, it seems appropriate to use fictional third-person stories as stimuli in the scanner. In a series of fMRI experiments, some researchers have therefore studied the neural basis of person perception by manipulating the contents of such verbal narratives (Gallagher et al., 2000; Vogeley et al., 2001; Saxe and Kanwisher, 2003; Saxe and Powel, in press). Full consideration of the methodological advantages, restrictions and necessary controls involved in this research is beyond the scope of this review; but, of course, the most persuasive reason of all to use verbal stimuli is that, empirically, experiments using these stimuli have produced robust and theoretically interesting results, some of which I will consider next.

4. How to use fMRI data: functional regions of interest

Once a hypothesis is identified and the stimuli are chosen, a third hurdle faces the social cognitive neuroscientist: how to understand the results. At its best, cognitive neuroscience can provide neural evidence to distinguish between competing theories of a cognitive function. In practice, data from fMRI

experiments can be notoriously difficult to interpret. Psychological theories concern information processing operations, whereas neuroimaging data can only test the change in blood oxygenation measured in a neural "voxel". What is the relationship between the two? It is critical to most cognitive neuroscience that, at least in some cases, anatomical divisions in the brain correspond to functional divisions in the mind and that these anatomical-functional units are consistently localised across individuals. But, even good correspondence between a patch of cortex and an interesting cognitive function is not sufficient for theoretical progress. Neuroimaging studies are often accused of discovering only "where" some psychological function is implemented in the brain, and since we already knew that cognitive functions occur in the brain, this is not big news. The challenge is to go beyond "where" to "what" and "how".

The first step toward understanding the neural basis of a cognitive function, like Theory of Mind, is to identify brain regions that are involved in the operation of that function. Hypotheses may come from lesion studies (e.g. Broca's and Wernicke's areas), from direct stimulation of the cortex (e.g. motor cortex) or from animal models (e.g. V1). For higher cognitive functions, many hypothesised region–function links come from simple, early imaging studies, which used broad task contrasts and whole brain analyses in individuals and groups to identify the implicated region(s) (e.g. for Theory of Mind, Gallagher et al., 2000). Not every pair of tasks differentially recruits a robust, reliable profile of regions: viewing faces (minus activity when viewing objects) does recruit a distinctive set of brain regions that can be identified in most individual subjects, but viewing familiar versus unfamiliar faces, for example, does not (Kanwisher and Yovel, in press). Broad contrasts can identify candidate regions for "where" a component of a psychological function occurs in the brain.

In order to test "how" psychological functions are neurally implemented, cognitive neuroscientists must then formulate and test hypotheses about the specific role that each of these candidate regions plays in the cognitive operation. One contribution that fMRI data may make to psychological theorising, for example, is the discovery that two similar tasks recruit distinct brain regions (and so, by hypothesis, distinct information processing components) or that two different tasks recruit a common brain region (and a common information processing component). Caution is required, though, when making either of these claims (for further discussion, see Brett et al., 2002; Saxe et al., 2004b, in press).

Consider, for example, the cortex near the right posterior superior temporal sulcus (R pSTS) extending into the right temporo-parietal junction (RTPJ). Early neuroimaging studies reported activations in these and neighbouring regions when subjects viewed human faces and facial expressions, eye gaze shifts, hand and body motions, and cartoons depicting 'animate motion,' as well as when subjects read short verbal stories about a protagonist's false beliefs (Allison et al., 2000). All of these stimulus conditions share some general "social" component but conflate a distinction that is evident in the trajectory of Theory of Mind development. Infants and toddlers use gaze as a cue to an actor's intentions and interpret hand actions as goal-directed (Woodward, 1998), but only around their fourth birthday do children first understand the specific (representational) contents of mental states like (false) beliefs (Wellman and Cross, 2001).

Based on the developmental evidence, we might therefore hypothesise that attribution of thoughts and beliefs—characteristic of the late-developing component of ToM—recruits a brain region near the R pSTS that is distinct from the region recruited by the (earlier developing) attribution of other internal states. How can we test this hypothesis?

It is not sufficient to note that enhanced BOLD signal has been reported near the R pSTS both in studies using intentional action stimuli and in studies using false belief stories and to therefore conclude that no distinct region exists for attributing representational mental states. "Right pSTS" refers imprecisely to a large swath of cortical real estate spanning 10 or more cm^2, probably encompassing many functionally distinct regions. (Distinct regions of extrastriate visual cortex, like MT, for example, are no more than one or 2 cm^2 in size.) One prevalent approach is therefore to compare the locations of observed "activations" in a "standard" brain, like the Talairach and Tournoux stereotaxic system (1988). The problem is that even well-known brain areas like V1 and Broca's area do not land consistently in the same place in the "normalised" brain from individual to individual (Amunts et al., 2000). Such variability across individuals is likely to blur regional boundaries in group data and lead to false conflations of neighbouring, but distinct, functional areas.

The best solution is to identify the regions of interest (ROIs)—that is, the regions that figure in the hypothesis—in each subject's brain independently and in advance of testing that hypothesis. In some cases, such identification can proceed based on anatomical markers, as for the amygdala, the calcarine sulcus and the precentral gyrus. More commonly, though, we lack physical markers to distinguish specific cortical areas in vivo. An alternative, then, is to use functional signatures to identify and distinguish cortical regions. Researchers studying early visual cortex, for example, use retinotopic mapping to identify the boundaries between V1 and V2, the foveal confluence, and so on. Category-selective regions of extrastriate cortex can be identified by their profile of response when the subject views faces, objects, bodies and scenes. Primary motor cortex can be identified by having the subject alternately move and then lie still.

The same logic can also be applied to social cognitive neuroscience. The only requirement is a contrast of tasks or of stimuli that (in principle) isolates an interesting aspect of person perception and that (in practice) leads to a selective, reliable and robust pattern of activation in most individual subjects. For example, Saxe and Kanwisher (2003) presented verbal stories modelled on the False Belief and False Photograph tasks used by developmental psychologists (Zaitchik, 1990). Twenty-eight subjects in the scanner read 24 short narratives about a representation (12 about a belief, 12 about a photo, drawing or map) that did not correspond to reality usually because the content of the representation became outdated and then answered a question either about the representation or about reality. In almost every subject, a very stereotyped set of brain regions showed higher BOLD recruitment during 'Belief' stories than 'Photo' stories: right and left

temporo-parietal junction, posterior cingulate, medial prefrontal cortex and right anterior superior temporal sulcus. The same contrast has now been used in our laboratory to identify ROIs in over a hundred individual brains.

These ROIs can then be used to test the hypothesis that attributing thoughts and beliefs recruits a distinct set of brain regions from the attribution of other (early-developing) internal states. For example, Saxe et al., (2004a) found that neighbouring but distinct regions were recruited by reading stories about beliefs (RTPJ) and when watching movies of intentional actions (right pSTS, Saxe et al., 2004a). Saxe and Powel (in press) found converging evidence for the same hypothesis. Subjects read stories from three conditions, each highlighting a different aspect of reasoning about another person: (1) 'Appearance' stories required detecting the presence of another person and representing socially relevant information about that person; (2) 'Bodily Sensations' stories elicited attribution of invisible, subjective and/or internal states, characteristic of early-developing components of Theory of Mind; and (3) 'Thoughts' stories described the contents of another person's thoughts or beliefs, specifically, the function of the later-developing component of Theory of Mind. The BOLD response in the RTPJ was high selectively when subjects read stories about a protagonist's thoughts or beliefs, but not when they read about subjective, internal physical feelings or other socially relevant information, such as appearance and personality attributes (see also Saxe and Wexler, 2005). These data converge on the conclusion that the cortex near the right pSTS includes at least two functionally dissociable regions: the RTPJ, which is involved in the late-developing component of Theory of Mind, is distinct from the neural correlate(s) of the early-developing component(s). Such precision is made possible through the use of individually tailored ROIs.

There are both methodological and statistical advantages to the use of ROIs (see also Saxe et al., in press). Using a predefined region of interest allows the researcher to test hypotheses in a statistically unbiased data set and to test new hypotheses about the *same* region as identified in previous experiments. The task used as the "localiser" experiment is often just a simple blocked contrast between two conditions, for maximum sensitivity, while the second experiment, analysed in the ROIs, can have a much more complicated design: multiple conditions, event-related habituation experiments, and so on. ROI analyses can also provide robust evidence for no difference between two conditions (e.g. Jiang et al., 2004).

The most common concern about the use of ROI analyses is that they will obscure the researchers' view of the bigger picture. A small but significant effect observed in the average response of an ROI might reflect the specific engagement of the region of interest, but it could also reflect a small but consistent effect that is actually occurring across many regions of the brain or the fringe or tail of a much bigger activation in the same contrast in a neighbouring region, or a different region, sending feedback connections to the region of interest. The solution is to combine ROI analyses with each other and/or with voxel-based whole brain analyses. Multiple simultaneous ROI analyses allow the researcher to test for significant differences between different brain regions' response profiles (e.g. Saxe et al., in press), and voxel-based whole brain tests how restricted those profile is to the preidentified regions (Saxe and Wexler, 2005).

In summary, ROIs are useful for specifying brain locations across subjects, for testing hypotheses concerning the function of specific brain regions and—occasionally—for investigating candidate separable components of the mind (Saxe et al., in press). The same benefits that make functional regions of interest ubiquitous in studies of retinotopic visual areas are equally important for theoretical progress in social cognitive neuroscience.

5. Conclusions

Obviously, the approaches described here are not the only, nor always the best, ways to approach the neuroscientific investigation of person perception. Developmental psychology is only one of the neighbouring disciplines with which social cognitive neuroscientists can and do have fruitful cross-fertilisations. Verbal narratives are stimuli well suited to the investigation of reasoning about representational mental states but would not be useful for an investigation of perceived gaze shifts. Regions of interest should be supplemented with whole brain analyses, and fMRI in general should be supplemented with other methods, including EEG and lesion studies. Still, the approaches described in this review have made, and will continue to make, important contributions in the study of Theory of Mind with fMRI.

Acknowledgment

Many thanks to Anna Jenkins for help with the manuscript.

REFERENCES

Allison, T., Puce, A., McCarthy, G., 2000. Social perception from visual cues: role of the STS region. Trends Cogn. Sci. 4, 267–278.

Amunts, K., Malikovic, A., Mohlberg, H., Schormann, T., Zilles, K., 2000. Brodmann's Areas 17 and 18 brought into stereotaxic space—Where and how variable? NeuroImage 11, 66–84.

Astington, J.W., Baird, J.A., 2005. Why Language Matters for Theory of Mind. Oxford Univ. Press, Oxford.

Astington, J.W., Jenkins, J.M., 1999. A longitudinal study of the relation between language and theory of mind development. Dev. Psychol. 35, 1311–1320.

Baldwin, D.A., 1993. Infants' ability to consult the speaker for clues to word reference. J. Child Lang. 20, 395–418.

Baron-Cohen, S., 1994. How to build a baby that can read minds: cognitive mechanisms in mindreading. Cah. Psychol. 13, 513–552.

Baumeister, R.F., Zhang, L., Vohs, K.D., 2004. Gossip as cultural learning. Rev. Gen. Psychol. 8, 111–121.

Baxter, L.A., Dun, T., Sahlstein, E., 2001. Rules for relating communicated among social network members. J. Soc. Pers. Relat. 18 (2), 173–199.

Bettelheim, B., 1975. The Uses of Enchantment. Random House, New York.

Bloom, P., 2000. How Children Learn the Meanings of Words. The MIT Press, Cambridge, MA. 300 pp.
Bloom, P., German, T.P., 2000. Two reasons to abandon the false belief task as a test of theory of mind. Cognition 77, B25–B31.
Brett, M., Johnsrude, I.S., Owen, A.M., 2002. The problem of functional localization in the human brain. Nat. Rev., Neurosci. 3 (3), 243–249.
Carlson, S.M., Moses, L.J., 2001. Individual differences in inhibitory control and children's theory of mind. Child Dev. 72, 1032–1053.
Carey, S., Spelke, E., 1994. Domain-specific knowledge and conceptual change. In: Gelman, L., Hirschfeld, S. (Eds.), Mapping the Mind: Domain Specificity in Cognition and Culture. Cambridge Univ. Press, New York, NY.
Carpenter, M., et al., 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monogr. Soc. Res. Child Dev. 63.
Carruthers, P., Smith, P.K. (Eds.), 1996. Theories of Theories of Mind. Cambridge Univ. Press, Cambridge. 390 pp.
Chomsky, N., 1980. Rules and representations. Behav. Brain Sci. 3, 1–61.
de Villiers, J., 2000. Language and Theory of Mind: what are the developmental relationships? In: Baron-Cohen, S., Tager-Flusberg, H., Cohen, D.J. (Eds.), Understanding Other Minds. Oxford Univ. Press, Oxford.
de Villiers, J., 2005a. Can language acquisition give children a point of view? In: Astington, J.W., Baird, J.A. (Eds.), Why Language Matters for Theory of Mind. Oxford Univ. Press, Oxford, pp. 86–219.
de Villiers, P., 2005b. The role of language in theory-of-mind development: what deaf children tell us. In: Astington, J.W., Baird, J.A. (Eds.), Why Language Matters for Theory of Mind. Oxford Univ. Press, Oxford, pp. 266–297.
Decety, J., Grezes, J., 1999. Neural mechanisms subserving the perception of human actions. Trends Cogn. Sci. 3, 172–178.
Dunbar, R.I.M., 2004. Gossip in evolutionary perspective. Rev. Gen. Psychol. 8 (2), 100–110.
Dunn, J., Brophy, M., 2005. Communication, relationships, and individual differences in children's understanding of mind. In: Astington, J.W., Baird, J.A. (Eds.), Why Language Matters for Theory of Mind. Oxford Univ. Press, Oxford, pp. 50–69.
Emler, N., 1994. Gossip, reputation and social adaption. In: Goodman, R.F., Ben-Ze'ev, A. (Eds.), Good Gossip. University Press of Kansas, Lawrence, pp. 119–140.
Figueras-Costa, B., Harris, PL., 2001. Theory of mind development in deaf children: a nonverbal test of false-belief understanding. J. Deaf Stud. Deaf Educ. 6, 92–102.
Flanagan, O., 1998. Ethics naturalized: ethics as human ecology. In: May, L., Friedman, M., Clark, A. (Eds.), Mind and Morals. MIT Press, Cambridge, MA.
Flavell, J.H., 1988. The development of children's knowledge about the mind: from cognitive connections to mental representations. In: Astington, J.W., Harris, P.L., Olson, D.R. (Eds.), Developing Theories of Mind. Cambridge Univ. Press, New York.
Fodor, 1983. The Modularity of Mind. MIT Press, Cambridge, MA.
Gallagher, H.L., Happe, F., Brunswick, N., Fletcher, P.C., Frith, U., Frith, C.D., 2000. Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. Neuropsychologia 38, 11–21.
Gallese, V., Goldman, A., 1998. Mirror neurons and the simulation theory of mindreading. Trends Cogn. Sci. 2, 493–501.
Gallese, V., Keysers, C., Rizzolatti, G., 2004. A unifying view of the basis of social cognition. Trends Cogn. Sci. 8, 396–403.
Goldman, A., 1993. Ethics and cognitive science. Ethics 103 (2), 337–360.
Gopnik, A., Astington, JW., 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance–reality distinction. Child Dev. 59, 26–37.
Gopnik, A., Wellman, H.M., 1992. Why the child's theory of mind really is a theory. Mind Lang. 7 (1–2), 145–171.
Harris, P.L., 1992. From simulation to folk psychology: the case for development. Mind Lang. 7, 120–144.
Harris, P.L., 2005. Conversation, pretense and theory of mind. In: Astington, J.W., Baird, J.A. (Eds.), Why Language Matters for Theory of Mind. Oxford Univ. Press, Oxford, pp. 70–83.
Harris, P.L., Johnson, C., Hutton, D., Andrews, G., Cooke, T., 1989. Young children's theory of mind and emotion. Cogn. Emot. 3, 379–400.
Haviland, J.B., 1977. Gossip, Reputation and Knowledge in Zinacantan. University of Chicago Press, Chicago.
Humphrey, C., 1997. Exemplars and rules: aspects of the discourse of moralities in Mongolia. In: Howell, S. (Ed.), The Ethnography of Moralities. Routledge, London, pp. 25–47.
Jiang, Y., Saxe, R., Kanwisher, N., 2004. Functional magnetic resonance imaging provides new constraints on theories of the psychological refractory period. Psychol. Sci. 15 (6), 390–396.
Johnson, M., 1993. Moral Imagination: Implications of Cognitive Science for Ethics. University of Chicago Press, Chicago.
Kanwisher, N., Yovel, G., in press. The fusiform face area: a cortical region specialized for the perception of faces. Philos. Trans. R. Soc. London B.
Lagattuta, K.H., Wellman, H.M., 2001. Thinking about the past: early knowledge about links between prior experience, thinking, and emotion. Child Dev. 72, 82–102.
Leslie, A., 1994. A theory of ToMM, ToBy, and Agency: core architecture and domain specificity. In: Hirschfeld, L., Gelman, S. (Eds.), Mapping the Mind: Domain Specificity in Cognition and Culture. Cambridge Univ. Press, New York, pp. 119–148.
Lohman, H., Tomasello, M., 2003. The role of language in the development of false belief understanding: a training study. Child Dev. 74, 1130–1144.
Moses, L.J., 2001. Executive accounts of theory-of-mind development. Child Dev. 72, 688–690.
Nichols, S., Stich, S., Leslie, A., Klein, D., 1996. Varieties of off-line simulation. In: Carruthers, P., Smith, P.K. (Eds.), Theories of Theories of Mind. Cambridge Univ. Press, Cambridge.
Pelphrey, K.A., Singerman, J.D., Allison, T., McCarthy, G., 2003. Brain activation evoked by perception of gaze shifts: the influence of context. Neuropsychologia 41, 156–170.
Perner, J., 1993. Understanding the Representational Mind. The MIT Press, Cambridge, MA.
Perner, J., Zauner, P., Sprung, M., 2005. What does "that" have to do with point of view? Conflicting desires and "want" in German. In: Astington, J.W., Baird, J.A. (Eds.), Why Language Matters for Theory of Mind. Oxford Univ. Press, Oxford, pp. 220–244.
Peterson, C.C., Siegal, M., 1999. Representing inner worlds: Theory of Mind in autistic, deaf and normal hearing children. Psychol. Sci. 10, 26–29.
Phillips, A.T., Wellman, H.M., Spelke, E.S., 2002. Infants' ability to connect gaze and emotional expression to intentional action. Cognition 85, 53–78.
Puce, A., Allison, T., et al., 1998. Temporal cortex activation in humans viewing eye and mouth movements. J. Neurosci. 18 (6), 2188–2199.
Repacholi, B.M., Gopnik, A., 1997. Early reasoning about desires: evidence from 14- and 18-month-olds. Dev. Psychol. 33, 12–21.
Saxe, R., 2005. Against simulation: the argument from error. Trends Cogn. Sci. 9 (4), 174–179.
Saxe, R., in press. Temporo-parietal and medial prefrontal cortices support distinct, uniquely human components of social cognition. Curr. Opin. Neurobiol.
Saxe, R., Kanwisher, N., 2003. People thinking about thinking people: fMRI studies of Theory of Mind. NeuroImage 19 (4), 1835–1842.
Saxe, R., Powell, L., in press. It's the thought that counts: specific

brain regions for one component of Theory of Mind. Psychol. Sci.

Saxe, R., Wexler, A., 2005. Making sense of another mind: the role of the right temporo-parietal junction. Neuropsychologia 43 (10), 1391–1399.

Saxe, R., Carey, S., Kanwisher, N., 2004a. Understanding other minds: linking developmental psychology and functional neuroimaging. Annu. Rev. Psychol. 55, 87–124.

Saxe, R., Xiao, DK., Kovacs, G., Perrett, D.I., Kanwisher, N., 2004b. A region of right posterior superior temporal sulcus responds to observed intentional actions. Neuropsychologia 42 (11), 1435–1446.

Saxe, R., Jamal, N., Powell, L., 2006. My body or yours? The effect of visual perspective on cortical body representations. Cereb. Cortex. 16 (2), 178–182.

Saxe, R., Brett, M.M., Kanwisher, N., in press. Divide and conquer: a defense of functional localisers. Neroimage

Scholl, B., Leslie, A., 1999. Modularity, development and 'theory of mind'. Mind Lang. 14 (1), 131–153.

Talairach, P., Tournoux, J., 1988. A Stereotactic Coplanar Atlas of the Human Brain. Stuttgart Thieme.

Tardif, T., Wellman, HM., 2000. Acquisition of mental state language in Mandarin- and Cantonese-speaking children. Dev. Psychol. 36, 25–43.

Varley, R., Siegal, M., 2000. Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. Curr. Biol. 10, 723–726.

Varley, R., Siegal, M., Want, S.C., 2001. Severe impairment in grammar does not preclude theory of mind. Neurocase 7, 489–493.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., et al., 2001. Mind reading: neural mechanisms of theory of mind and self-perspective. NeuroImage 14, 170–181.

Wellman, H.M., Cross, D., 2001. Theory of mind and conceptual change. Child Dev. 72, 702–777.

Wellman, H.M., Woolley, J.D., 1990. From simple desires to ordinary beliefs: the early development of everyday psychology. Cognition 35, 245–275.

Wellman, H.M., Harris, P.L., Banerjee, M., Sinclair, A., 1995. Early understanding of emotion: evidence from natural language. Cognition and Emotion 9, 117–149.

Wellman, H.M., Cross, D., Watson, J., 2001. Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev. 72, 655–684.

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

Woodward, A.L., 1998. Infants selectively encode the goal object of an actor's reach. Cognition 69, 1–34.

Zaitchik, D., 1990. When representations conflict with reality: the preschooler's problem with false beliefs and "false" photographs. Cognition 35, 41–68.

Ziv, M., Frye, D., 2003. The relation between desire and false belief in children's theory of mind: no satisfaction? Dev. Psychol. 39 (5), 859–876.