Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects

Rebecca Saxe, Joseph M. Moran, Jonathan Scholz, and John Gabrieli
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02138, USA

INTRODUCTION

The classic task for assessing a child’s ability to reason about the mental states of others (her ‘theory of mind’) is the False Belief task (Wimmer and Perner, 1983; for reviews of this literature, see Flavell, 1999; Wellman et al., 2001). In the standard version of this task (the ‘object transfer’ problem), the child is told a story in which a character’s belief about the location of a target object becomes false when the object is moved without the character’s knowledge. The critical feature of a False Belief task is that to reach the correct answer, the child must pay attention to the character’s belief, and not just to the actual location of the object (Dennett, 1978). Dozens of versions of the False Belief problem have been used, and while the precise age of success varies between children and between task versions (Wellman et al., 2001); in general, children <3 or 4 years old do not correctly solve False Belief problems, but older children do.

Many neuroimaging studies have followed developmental psychology in using False Belief problems as the definitive Theory of Mind task (e.g. Fletcher et al., 1995; Gallagher et al., 2000; Vogeley et al., 2001; Ruby and Decety, 2003; Saxe and Kanwisher, 2003). These studies have revealed an impressively consistent pattern of brain regions involved when subjects are required to reason about someone’s false belief, including the medial prefrontal cortex (MPFC), medial prefrontal and bilateral temporo-parietal junction (left: LTPJ, right: RTPJ).

What is the distinct contribution of each of these regions to the subject’s reasoning about other people? A series of recent results suggest that while the RTPJ is recruited specifically when subjects think about a character’s thoughts, the medial prefrontal and MPFC are recruited more generally for many different judgements about people (Bermphol, 2004; Mitchell et al., 2005a, b; Saxe and Wexler, 2005; Northoff and Berghol, 2004; Saxe, 2006; Saxe and Powell, 2006). In particular, one line of research reliably reports higher response in both medial regions, but not in the lateral TPJ regions, when subjects judge whether a trait adjective applies to them (‘self task’), than when subjects make semantic judgements about the same adjectives (‘semantic task’, Gusnard et al., 2001; Kelley et al., 2002; Macrae et al., 2004; Schmitz et al., 2004; see also D’Argembeau et al., 2005; Goldberd et al., 2006; Northoff et al., 2006; Ochsner et al., 2006). These results may provide (i) hints about the distinct contributions of the medial and lateral components of the ‘theory of mind network’ in the brain, (ii) a rare functional dissociation between brain regions that often activate, deactivate and even spontaneously fluctuate together (Greicius et al., 2003) and (iii) impetus for cognitive psychologists to determine the common function underlying both ‘false belief’ and ‘self-trait’ attributions.

Each of these implications depends on a strong claim about the overlap between the ‘false belief’ and ‘self’ tasks in medial cortex: those similar-looking group activations are the result of recruitment in the very same regions of individual subjects for both tasks. However, group analyses in normalized brain-space produce blurred activation maps,
due to the necessarily imperfect registration across physically different brains. Individuals vary not only in their physical anatomy but also in their functional anatomy, producing yet more blurring in group-averaged data. Thus, activations that may be completely non-overlapping within each individual could be highly overlapping when the same data are averaged across subjects. This problem is exacerbated when comparing activations across subject groups or across studies. In the current study, we therefore directly compared the results of the ‘false belief’ and ‘self’ tasks in individual subjects.

**METHODS**

Eight naive, right-handed adults participated in the functional magnetic resonance imaging (fMRI) study for payment. All subjects were native English speakers, had normal or corrected-to-normal vision, and gave written informed consent in accordance with the requirements of internal review boards at MIT. Subjects were scanned using a Siemens Magnetom Tim Trio 3T system (Siemens Medical Solutions, Erlangen, Germany) in the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, using thirty 4-mm-thick near-axial slices covering the whole cortex. Standard echoplanar imaging procedures were used (TR = 2 s, TE = 30 ms, flip angle = 90°).

In the theory of mind experiment, subjects read 24 short narratives about the formation of a representation (12 about beliefs, 12 about physical representations like a photo, drawing or map) that did not correspond to reality (Saxe and Kanwisher, 2003). Stories were on average 32 words long, and were presented for 10 s. Subjects then answered a fill-in-the-blank question either about the representation or about reality (presented for 4 s). Stories from the two conditions alternated, with a 12-s rest period after each story. Each run lasted 2 min and 48 s (six stories); each subject participated in four runs of this experiment.

During the self-attribution experiment subjects viewed a series of 200 trait adjectives presented across five functional runs. Words were drawn from Anderson’s (1968) list of normed trait adjectives, and lists were counterbalanced for word valence, length and number of syllables. Words were presented in a blocked design. Each word was presented for 3 s in blocks of ten. Prior to each block onset subjects viewed a 2 s cue screen describing their task for the upcoming block. Subjects either judged the words in the following block for their self-descriptiveness (‘Does this word apply to you?’) or for their valence (‘Is this word positive?’). Order of conditions was counterbalanced within and across subjects, and each block was followed by 10 s of rest. Each run lasted for 3 min and 4 s.

Stories and words were projected onto a screen via Matlab 5.0 running on an Apple G4 laptop computer, in white 24-point font on a black background. Because of technical errors, behavioral data were not collected in the scanner. However, extensive behavioral data on these stimuli are available from previous studies (e.g. Kelley et al., 2002; Saxe and Kanwisher, 2003; Macrae et al., 2004).

The fMRI data were analyzed with SPM2 and in-house software. Individual subjects’ data were motion corrected, and then smoothed using a Gaussian filter (full width half maximum = 5 mm), and high-pass filtered during analysis. Both fMRI experiments used a blocked design and were modeled using a boxcar regressor.

All analyses were conducted in individual subjects. Voxel weights were labeled as ‘overlapping’ if the t-value for each contrast (false belief > false photograph) and (self > semantic), was independently greater than 3.6 ($P < 0.001$, uncorrected). Voxel weights were labeled as ‘non-overlapping’ if the t-value for one contrast exceeded 3.6, and the t-value of the other contrast was below 0.5. This low threshold reduced the chance of false positives inflating the observed non-overlap.

Theory of Mind regions of interest in each subject were defined as clusters of contiguous voxels with a higher BOLD response during ‘false belief’ than ‘false photo’ stories ($P < 0.0001$, uncorrected), within 9 mm of the peak voxel in anatomical areas implicated in theory of mind by previous studies: precuneus, MPFC and bilateral TPJ. Using the same threshold, we defined ROIs in the precuneus and MPFC, based on the response to the ‘self task’ vs the ‘semantic task’. The percent signal change over each block, relative to rest, was then estimated in each ROI for both tasks.

**RESULTS**

The current design allowed us to ask two questions about the relationship between brain regions recruited for the ‘false belief’ and ‘self’ tasks. First, are there sub-regions significantly recruited for both tasks (i.e. is there any real overlap)? Second, are there sub-regions recruited significantly for one task, and not recruited for the other task (i.e. is there any real non-overlap)?

One approach to these questions is to examine the overlap and non-overlap between whole brain contrast maps for each task, in each individual subject. Regions of overlap between both contrasts were observed in the precuneus in 8/8 subjects (Figure 1a), and in MPFC in 6/8 subjects (Figure 1b), although note that regions of non-overlap were also observed. For each anatomical region of interest, for each subject, we calculated the total number of voxels recruited ($P < 0.001$) for either task. We then calculated the proportion of this total activation that could be conservatively classified as overlapping (i.e. $P < 0.001$ in both tasks), or non-overlapping (i.e. $P < 0.001$ in one task, and $P > 0.3$ in the other task, Figure 2). This profile differed significantly by region of interest (region by functional classification interaction $F(6,36) = 3.4, P < 0.01$, no main effects). In the MPFC, voxels were most likely to be either recruited by both tasks or by the ‘self’ task only. In the precuneus, voxels were most likely to either recruited by both tasks or by the ‘theory of mind’ task only. In the TPJ bilaterally, voxels were most likely to be recruited only by the ‘theory of mind’ task.
Another approach is to define functional regions of interest based on one contrast, and then evaluate the response in the same voxels, in the other task. Based on the theory of mind experiment, we found ROIs in the RTPJ (8/8 subjects), medial precuneus (8/8), MPFC (6/8) and LTPJ (4/8, too few for further analyses). In ROIs identified based on the theory of mind task, the response was higher during ‘self’ than ‘semantic’ judgements in the medial precuneus \( t(7) = 3.1, P < 0.02 \) and MPFC \( t(5) = 4.5, P < 0.01 \), but not in the RTPJ \( t(7) = 1.3, \) NS, Figure 3]. Based on the self task, we found ROIs in the medial precuneus (8/8 subjects) and MPFC (7/8). In these ROIs, the response was higher during ‘false belief’ vs ‘false photograph’ stories \[ \) precuneus: \( t(7) = 2.9, P < 0.05 \); mpfc: \( t(6) = 4.6, P < 0.01 \), Figure 4].

**DISCUSSION**

The current results are consistent with previous group analyses. Deciding whether a trait adjective applies to oneself recruit the medial regions associated with theory of mind—the medial precuneus and MPFC—but not the TPJ.

Caution in interpreting the observed overlap is appropriate. These data do not completely overcome methodological limitations. First, a single voxel in the current study reflects the average response over 3 x 3 x 4 mm of tissue, much lower spatial resolution than the functional organization of cortex. Higher resolution functional imaging is currently becoming available, and may yet reveal that activations associated with the ‘self’ and ‘false belief’ tasks are neighboring but distinct (e.g. Schwarzlose et al., 2005). Second, even within single voxels, neurons subserving distinct functions may be interspersed. One approach to disentangling such overlap is functional adaptation, which relies on the reduction of activity observed when two successive stimuli are processed by the same sub-population of neurons within a voxel, but not when the stimuli recruit different sub-populations (Kourtzi and Kanwisher, 2001;
Krekelberg et al., 2006). Third, although overlapping voxels were observed in each individual, in most individuals we also observed non-overlapping voxels in medial regions, in which the t-value of one task was higher than 3.6, and the second task did not reach \( t > 0.5 \) (corresponding to a \( P \)-value of 0.3). These non-overlapping voxels provide evidence that at least some aspects of the medial regions’ contribution to theory of mind are not shared by the self-attribution task and vice versa.

Still, the current data provide the strongest evidence to date that sub-regions of medial precuneus and MPFC are recruited both when subjects reason about a character’s thoughts, and when they attribute a personality trait to themselves. Recently, research has revealed a third cognitive function associated with very similar regions of medial cortex: autobiographical episodic memory (e.g. Shannon and Buckner, 2004; Wheeler and Buckner, 2004; Wagner et al., 2005; Ries et al., 2006; see also Fossati et al., 2004; Lou et al., 2004).

Interestingly, these same three tasks—theory of mind, self reflection and autobiographical episodic memory—are correlated in child development (Moore and Lemmon, 2001). One measure of self-reflection in childhood is Povinelli and colleagues’ (1996) delayed self-recognition task. In this task, an experimenter is videotaped covertly placing a large sticker on the child’s head. Three minutes later, the child is shown the video tape. Although all children between 2 and 4 years correctly identify themselves in the video, only children over 3.5 years reach up to retrieve the sticker. Performance on this task specifically reflects children’s developing conception of the connection between their past and present selves; given a mirror, children at all these ages successfully retrieve the sticker. Children’s performance on the delayed-self recognition task is correlated with scores on episodic memory and false belief tasks (Moore and Lemmon, 2001).

Developmental and neural data thus converge on a triad of interrelated tasks. The next challenge is to establish the causal and dependence relations between these tasks, and/or the common cognitive function(s) underlying them. Both careful studies of individual subjects’ functional data, and careful task analyses, will be necessary. Many different models are consistent with the current evidence (Figure 5):

1. Autobiographical memory may depend on theory of mind. Perner (2001) has proposed that autobiographical/episodic memory inherently depends on theory of mind: that is, on the ability to identify the source of a current experience (a recollection) in the previously experienced event. Developmental evidence is consistent with such a dependence of autobiographical memory on theory of mind. Children’s theory of the origin of epistemic states—that is, of how beliefs and knowledge are acquired or caused—develop along with performance on false belief tasks (Wimmer et al., 1988; O’Neill et al., 1992). Three year olds, for example, but not four year olds, expect that people (including themselves) can distinguish between a heavy ball and a light ball just as well by looking at the balls as by lifting them (Burr and Hofer, 2002). Success on tests of theory of mind predicts success on episodic memory tasks (Perner, 2001), and helps children become resistant to contamination of their autobiographical memories through suggestion (Welch-Ross, 1999).
2. Theory of mind may depend on autobiographical memory. Some theorists have suggested that in order to understand the causal relations between another person’s experiences, thoughts, and behaviors, observers bring to mind specific, and relevantly similar, past experiences of their own (Adams, 2001). There is at least some evidence that empathy is affected by the observer’s prior experiences (Batson et al., 1996).
3. Self-reflection may depend on autobiographical memory. During the self-attribution task, subjects are asked to judge whether trait words (‘logical’, ‘reckless’, ‘rebellious’) apply to themselves. To answer this question, subjects may retrieve autobiographical memories of specific incidents in which their actions merited (or did not merit) the target description.
(4) Autobiographical memory and self-reflection may both depend on recognizing the self as an enduring entity, with persisting causal and social properties (Povinelli and Simon, 1998). Povinelli (2001) therefore proposed that delayed self-recognition is a necessary precursor to autobiographical memory.

(5) Theory of mind may depend on self-reflection. ‘Simulation Theory’ proposes that an observer attributes mental states to another person by using her own mind as a model of the other mind (e.g. Gallese and Goldman, 2004; but see Saxe, 2006). The observer would adjust (i) the input, using the other person’s (hypothized) perceptual environment, rather than her own and (ii) the output, generating a prediction rather than an action (Nichols and Stich, 2003). Identifying the output as a prediction for someone else’s action might involve a kind of self-attribution, and be a necessary component of theory of mind.

(6) Theory of mind and self-trait attribution may share a common conception of human agents as enduring entities, with persisting causal and social properties. This distinction lies at the core of recent proposals that there is a general domain of ‘social cognition’, distinct from all forms of ‘non-social cognition’ (Jenkins and Mitchell, in press; Mitchell et al., 2005b).

One further observation may illuminate (or complicate) this picture: patients with Alzheimer’s disease show amyloid deposition, hypo-metabolism, hypo-activation and tissue atrophy in these midline regions (Greicius et al., 2004; Shannon and Buckner, 2004; Buckner et al., 2005; Rombouts et al., 2005; Wang et al., 2006). This pattern converges with the evidence that medial precuneus and MPFC are involved in autobiographical memory, which is impaired in Alzheimer’s disease. However, performance on false belief tasks is preserved in patients with Alzheimer’s disease (Gregory et al., 2002; Zaitchik et al., 2004; Zaitchik et al., 2006), and there is also evidence for preserved self-attribution of traits (Klein et al., 2003; Cotrell and Hooker, 2005; Rankin et al., 2005). In contrast, both theory of mind and self-attribution task performance is impaired in a different degenerative disorder, fronto-temporal dementia (Gregory et al., 2002; Rankin et al., 2005). One study recently directly compared recruitment of the medial precuneus regions for autobiographical memory and for self-trait attribution in healthy subjects and in patients with mild cognitive impairment (MCI), a risk factor for Alzheimer’s disease. MCI patient showed hypo-activation of the medial precuneus and MPFC for the memory task but normal activation for the self-trait attribution (Ries et al., 2006). Thus while theory of mind, self-attribution and episodic memory are correlated in development, and recruit common brain regions in healthy adults, the three tasks appear to be dissociable in degenerative disease. Any full account of the role of the medial precuneus and MPFC should aim to explain all of these results.

Conflict of Interest
None declared.

REFERENCES