ELSEVIER

Neuropsychologia xxx (2005) xxx–xxx

NEUROPSYCHOLOGIA

www.elsevier.com/locate/neuropsychologia

Rapid publication

Making sense of another mind: The role of the right temporo-parietal junction

Rebecca Saxe [a,b,*], Anna Wexler [a]

[a] *Department of Brain and Cognitive Science, MIT, NE20-443, 77 Massachussetts Avenue, Cambridge, MA 02139, USA*
[b] *Department of Psychology, Harvard University*

Received 30 July 2004; received in revised form 21 January 2005; accepted 28 February 2005

Abstract

Human adults conceive of one another as beings with minds, and attribute to one another mental states like perceptions, desires and beliefs. That is, we understand other people using a 'Theory of Mind'. The current study investigated the contributions of four brain regions to Theory of Mind reasoning. The right temporo-parietal junction (RTPJ) was recruited selectively for the attribution of mental states, and not for other socially relevant facts about a person, and the response of the RTPJ was modulated by the congruence or incongruence of multiple relevant facts about the target's mind. None of the other three brain regions commonly implicated in Theory of Mind reasoning – the left temporo-parietal junction (LTPJ), posterior cingulate (PC) and medial prefrontal cortex (MPFC) – showed an equally selective profile of response. The implications of these results for an alternative theory of reasoning about other minds – Simulation Theory – are discussed.

Keywords: Theory of Mind; Simulation; Attribution; Social cognitive neuroscience

1. Introduction

By a 'Theory of Mind', we mean the process(es) by which most healthy human adults (1) attribute unobservable mental states to others (and under certain circumstances, to the self, [cf. Bem, 1967; Happe, 2003]), and (2) integrate these attributed states into a single coherent model (Gopnik & Meltzoff, 1997) that can be used to explain and predict the target's behaviour and experiences. In this paper, we show that the hemodynamic response of one brain region – the right temporo-parietal junction (RTPJ) – reflects both of these characteristics of a Theory of Mind. First, we find that enhanced BOLD response in this region is selective to the attribution of mental states, and is not recruited by processing other socially relevant facts about a person. Second, activity in the RTPJ is modulated by the congruence or incongruence of multiple relevant facts about the target's mind. RTPJ activity was enhanced when the protagonist of a story professed a belief or desire that was inconsistent with the subject's expectations, based on the protagonist's background. Finally, none of the other brain regions commonly implicated in Theory of Mind reasoning – the left temporo-parietal junction (LTPJ), posterior cingulate (PC) and medial prefrontal cortex (MPFC) – showed an equally selective profile of response.

Recent neuroimaging work has suggested that multiple regions of cortex in the human brain are dedicated to components of the process of perceiving and reasoning about other people, including recognising and identifying human faces (e.g. Kanwisher, McDermott, & Chun, 1997; Hoffman & Haxby, 2000; Grill-Spector, Knouf, & Kanwisher, 2004), perceiving other human bodies (e.g. Downing, Jiang, & Kanwisher, 2001; Saxe, Jamal, & Powell, 2005), identifying human-like biological motion (e.g. Vaina et al., 2001; Grossman & Blake, 2002; Beauchamp, Lee, Haxby, & Martin, 2003; Pelphrey, Singerman, Allison, & McCarthy, 2003), perceiving intentional actions (e.g. Castelli, Happe, Frith, & Frith, 2000; Schultz et al., 2003; Saxe, Xiao, Kovacs, & Perret, 2004), and orienting towards and recognising basic emotional expressions (e.g. Whalen et al., 2001;

* Corresponding author. Tel.: +1 617 425 3127; fax: +1 617 258 8654.
E-mail address: saxe@mit.edu (R. Saxe).

0028-3932/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2005.02.013

LaBar, Crupain, Voyvodic, & McCarthy, 2003; Winston, O'Doherty, & Dolan, 2003; Wicker et al., 2003; Pessoa & Ungerleider, 2004). Beyond perceiving the physical appearance and behaviours of others, though, we intuitively conceive of each person as a being with a mind, and attribute to one another specific, content-ful mental states like perceptions, desires and beliefs. That is, we understand other people using a 'Theory of Mind' (Premack and Woodruff, 1978).

Could the human brain contain one or more specialised neural substrate(s) for Theory of Mind (e.g. a 'Theory of Mind Module' (Leslie & Thaiss, 1992)? At least four cortical regions are consistently identified as possible candidates (Saxe & Kanwisher, 2003; Gallagher et al., 2000; Fletcher et al., 1995, see reviews by Frith & Frith, 2003; Saxe, Carey, & Kanwisher, 2004): the right and left temporo-parietal junctions (RTPJ and LTPJ), posterior cingulate (PC) and medial prefrontal cortex (MPFC). The cognitive neuroscience of Theory of Mind has mostly depended on adaptations of the False Belief paradigm from developmental psychology. In this task, subjects must predict a character's action based on the character's false belief (Wimmer & Perner, 1983). False beliefs provide a useful behavioural test of Theory of Mind, because when the character's belief is false, the action predicted by the belief is different from the action that would be predicted by the true state of affairs (Dennett, 1978).

Increased activation has been reported in the same four brain regions when subjects reason about the false belief either of a character in a story (e.g. Fletcher et al., 1995; Gallagher et al., 2000; Vogeley et al., 2001) or in a cartoon (Gallagher et al., 2000; Brunet et al., 2000), or an imaginary ill-informed protagonist (Goel, Grafman, Sadato, & Hallett, 1995, a medical lay person, Ruby & Decety, 2003), relative to when belief attribution is not required. The control conditions have included scrambled texts, scrambled pictures (Fletcher et al., 1995; Gallagher et al., 2000), texts describing logical relations between events (Vogeley et al., 2001), judgements about the true function of an object (Goel et al., 1995), and stories about a physical representation of the world (e.g. a photograph) that becomes false (Saxe & Kanwisher, 2003, Experiment 2).

However, False Belief tasks alone are not enough to establish that a brain region is selectively recruited for Theory of Mind (e.g. Scholl & Leslie, 2001; Saxe, Carey et al., 2004), let alone to determine which component of Theory of Mind is reflected in its response. The current study was designed to help characterise the contributions of the RTPJ, LTPJ, PC, and MPFC to perceiving and understanding other people.

First, we asked whether the RTPJ, LTPJ, PC and/or MPFC was recruited selectively for the attribution of mental states, or more broadly when subjects reasoned about any socially relevant information about a person. A previous study (Saxe & Kanwisher, 2003) found that these regions were not involved in representing the mere physical appearance of another person. In the current study, we extended these earlier results by comparing the neural response during the attribution of mental states (e.g. 'he wants to be a neurosurgeon', 'she likes to watch a little TV', 'he thinks it is a good idea to have sex before marriage'), with the response to information about a protagonist's social, geographical or cultural background (e.g. 'Kevin is from Ireland and was raised strictly Catholic', 'Olivia comes from a middle class family', 'Carla has a top position at a large company'). The background information allowed subjects to begin to form an impression of the protagonist, without containing an explicit description of her mental states. Brain regions recruited selectively for mental state attribution should therefore respond little to the social background of the protagonist, while brain regions involved more broadly in social cognition and person perception might be recruited equally for both background and mental state information.

In addition, we asked whether the neural response would be affected by a manipulation of the protagonist's background: half of the protagonists had the same kinds of backgrounds as our subjects (the 'Familiar' backgrounds, e.g. middle class, American, urban), while the other half had 'Foreign' backgrounds (e.g. aristocratic, orthodox, isolated; see Appendix A for examples). We reasoned that the 'Foreign' background would constitute relevant distinctive social information about a person, in the absence of mental state attribution, and therefore would produce an enhanced response in brain regions involved in person perception but not restricted to Theory of Mind. For any brain region that was truly selectively involved in reasoning about mental states, we predicted (1) a low overall response to background information, and (2) no difference in the response to 'Familiar' versus 'Foreign' backgrounds.

Second, we asked whether the recruitment of each brain region would be influenced by the effort required to create an integrated coherent model of the protagonist's mind. Although tasks tapping Theory of Mind often measure the attribution of specific individual beliefs (or belief–desire pairs, e.g. Wimmer & Perner, 1983; Repacholi & Gopnik, 1997), mental state attribution is fundamentally holistic: a belief or desire can only be used to explain an action against the background of many (probably infinitely many) other beliefs and desires. Even 2- and 3-year-olds, in their spontaneous speech about the mind, obey the rule that mentioned mental states must be consistent with one another, and relevant to the action or situation (Bartsch & Wellman, 1995). Furthermore, there is extensive evidence that perceivers expect other people to be coherent, unified entities, and strive to resolve inconsistencies with that expectation (see review by Hamilton & Sherman, 1996). Consequently, when a target's behaviour violates the perceiver's previous impression of that person, the perceiver spends more time processing the behaviour (Bargh & Thein, 1985) and searching for the behaviour's causes (Hamilton, 1988), and later shows enhanced memory for the incongruent information (e.g. Wyer & Gordon, 1982, see review by Higgins & Bargh, 1987). We hypothesised that a similar process of integration and inconsistency-resolution could be elicited in the context of mental state attribution.

Expectations about the background mental states of a target individual may be influenced by schematic knowledge about the group membership of the target, and about the typical beliefs and desires of members of that group. We sought to set up such expectations about the protagonists of our stories in the background information, described above. Following the background information, we gave subjects a description of the protagonist's beliefs and/or desires. This mental state could be 'Normal' (similar to those of our subjects) or 'Norm-Violating' (unusual, and even inappropriate, in our subjects' social environment). The experiment followed a 2×2 design: the protagonist's mental state could either be congruent with her background (e.g. a 'Normal' mental state in a protagonist from a 'Familiar' background), or incongruent with her background (e.g. a 'Norm-violating' mental state in a protagonist from a 'Familiar' background, Terwogt & Rieffe, 2003).

After the mental state was described, the stories concluded with the outcome for the protagonist – whether her preference was fulfilled – and then subjects were asked to predict whether the protagonist would feel positive or negative about this outcome. Successful performance of the task depended on integration of the stated mental state of the protagonist with the subsequent outcome, but did not involve the protagonist's background. Subjects could therefore have adopted a policy of ignoring the background information altogether. By contrast, we anticipated that subjects would attempt to integrate all of the information about the protagonist's mind as it became available. Consequently, we predicted that the incongruent stories would elicit an effort to resolve the inconsistency, at the time when the mental state information was presented, and that this inconsistency resolution would be reflected in the BOLD response of brain regions involved in Theory of Mind.

These same stimuli also allowed us to test a third hypothesis about the way perceivers reason about other minds: namely, that the mind of the target is represented fundamentally in terms of the similarity between the target's mind, and the perceiver's own. This alternative hypothesis can be derived from one currently popular class of theories of understanding other minds, collectively called the Simulation Theory (ST, Stich & Nichols, 1992; Nichols, Stich, Leslie, & Klein, 1995). ST proposes that an observer reasons about other minds by 'putting herself in the other person's shoes' and then passively reading off the mental states that arise in her own mind, within the pretend context.

There are many different specific versions of ST, and so there can be no monolithic prediction for neural activity from an ST perspective. We reasoned that one way to cash out the central notion of simulation would be to predict a linear relationship between the similarity of the modelled mind to the modellers mind, and the response of brain regions involved in the simulation. In our paradigm, the protagonists from a 'Familiar' background who professed a 'Normal' belief or desire were the most similar to the observers (our subjects). The protagonists from a 'Foreign' background who professed 'Norm-violating' beliefs were the least similar, since they differed from our subjects both in background and in mental state. The other two conditions were intermediate: each contained one similar and one dissimilar element.

These two hypotheses – one inspired by ToM and the other by ST – thus make different predictions about the response of the RTPJ, LTPJ, PC, and MPFC (or any neural substrate of reasoning about other minds). The ToM perspective predicts a higher response for the incongruent conditions, relative to the congruent conditions. On our reading, ST might predict that the response of these regions would increase linearly with the similarity between the subject and the protagonists, reflecting the ease or success of the simulation. Alternatively, ST might predict that the neural response would increase linearly with the dissimilarity between the subject and the protagonists, reflecting the effort of the simulation or the number of changes to the default assumption of similarity (Harris, 1992; Nichols et al., 1995). However, our simple first-order version of ST does not predict an interaction between the background and mental state of the character. In the current study, we tested these competing predictions.

2. Methods

Twelve naïve right-handed subjects (6 female; 1 Asian, 2 African-American, 1 Hispanic) gave written informed consent in accordance with the requirements of Internal Review Boards at Massachussetts General Hospital and MIT. All subjects were native speakers of English, and had normal or corrected-to-normal vision. Furthermore, all subjects were raised in middle class families in the United States (for more details, see Section 3.1).

Subjects were scanned at 3T (at the MGH scanning facility in Charlestown, Massachusetts) using 26 4-mm-thick near-axial slices covering the whole brain except for the cerebellum. Functional scans used: TR = 2 s; TE = 40.

Story stimuli were modelled after Terwogt and Rieffe (2003), and consisted of 8 different variations of 12 different story topics (e.g. monogamy, violence and arranged marriage) for a total of 96 stories with an average of 80 words per story. We used a $2 \times 2 \times 2$ design for each story topic. First, each protagonist was either from a 'Familiar' moderate Western background or a 'Foreign' background (in terms of geography, religion, wealth or politics, see Appendix A for examples). Second, s/he either had a 'Normal' desire or a 'Norm-violating' desire. The 'Normal' versus 'Norm-violating' mental states were defined from our subjects' perspective, not from the perspective of the protagonist's social group. Each 'Norm-violating' mental state was constructed to be compatible with (i.e. conventional from the perspective of) the 'Foreign' background with which it was paired. Finally, the protagonist either got what s/he wanted or did not get what s/he wanted.

Following the scan, in a brief survey we confirmed that subjects found the moderate Western backgrounds 'Familiar' and that they shared the 'Normal' desires. The survey

first asked 'Which of these groups describes you or your family? Rate from 1 (not at all) to 5 (perfectly).' Subjects rated three familiar and seven foreign backgrounds. Next, the survey asked 'How much do you agree with the following beliefs or desires? Rate from 1 (not at all) to 5 (perfectly).' Subjects rated five normal and nine norm-violating desires.

Stories were presented in a pseudo-random order, counterbalancing the order of story conditions across runs and across subjects, thereby ensuring that no condition was immediately repeated. Subjects saw two versions of each story topic, for a total of 24 stories. When a story topic was repeated, the repetition contained a different protagonist (i.e. first name), background, desire, and outcome from the first presentation. The text of the stories was presented in a white 18-point font on a black background.

Stories were presented in three sections. First, sentences describing the character's background were presented on the screen for 6.3 s. Then, sentences describing the character's desire were added onto the screen and displayed for another 6.3 s. Finally, sentences related to the outcome of the story were displayed for 7.4 s, so that the story was presented for a total of 20 s. The story was then removed from the screen and replaced with the probe question: 'How will X (the protagonist) feel about this outcome? Positive or Negative.' The words 'positive' and 'negative' were displayed at the left and right side of the screen (in counterbalanced order) and the subject pressed the left/right button on a button box to choose his/her response. The question remained on the screen for 4 s.

Twelve stories were presented in each run. Fixation blocks of 12 s were interleaved between each story. Each run lasted 444 s. Subjects saw two runs of this experiment. The same subjects were also scanned on a localiser experiment, contrasting stories that required inferences about a character's beliefs with stories about a physical representation (e.g. photograph or map) that became outdated. Stimuli and story presentation were exactly as described in (Saxe & Kanwisher, 2003, Experiment 2).

2.1. fMRI analysis

MRI data were analysed using SPM 99 (http://www.fil.ion.ucl.ac.uk/spm/spm99.html) and in-house software. Each subject's data were motion corrected and then normalised onto a common brain space (the MNI template). Data were then smoothed using a Gaussian filter (full width half maximum = 5 mm), and high-pass filtered during analysis. Every experiment used a blocked design, and was modelled using a boxcar regressor.

Four regions of interest (ROI) were defined for each subject individually based on a whole brain analysis of the localiser experiment, and defined as contiguous voxels that were significantly more active ($p < 0.0001$, uncorrected) while the subject read stories about beliefs than about photographs: right and left temporo-parietal junctions (RTPJ and LTPJ), medial prefrontal cortex (MPFC) and posterior cingulate (PC). All peak voxels are reported in MNI coordinates.

The responses of these regions of interest were then measured while subjects read stories from the current experiment. Within the ROI, the average percent signal change (PSC) relative to fixation baseline (PSC = 100 × raw BOLD magnitude for (condition − fixation)/raw BOLD magnitude for fixation) was calculated for each condition at each time point (averaging across all voxels in the ROI as well as all blocks of the same condition).

A separate PSC was calculated for two segments of the story: background (the first 6.3 s) and mental state (the next 6.3 s) corrected for hemodynamic lag. (Following Terwogt and Rieffe (2003), the results were collapsed across the dimension of outcome). These values were then entered into repeated measures ANOVAs.

Because the data defining the ROIs were independent from the data used in the repeated measures statistics, Type I errors were drastically reduced.

3. Results

3.1. Behavioural results

The average familiarity score for familiar backgrounds was 4.32 out of a maximum of 5 (S.D. 0.72). For foreign backgrounds, the average score was 1.42 (S.D. 0.31, $t(1, 10) = 12.14$, $p < 0.001$, paired-samples t-test). The average agreement subjects reported with normal desires was 4.16 (S.D. 0.66) and for norm-violating desires it was 1.55 (S.D. 0.35, $t(1, 10) = 11.30$, $p < 0.001$, paired-samples t-test). No subject reported strong (>3) identification with any 'foreign' background or 'norm-violating' belief. These results confirm that our manipulation of 'Familiar' versus 'Foreign' backgrounds was valid for our subjects.

Behavioural data was collected from subjects in the scanners (behavioural data for two subjects was lost due to technical difficulties). Two-way ANOVAs (background by desire) of reaction times on correct trials, and of percent correct over all, revealed no main effects or interactions. Reaction times for the four conditions were: familiar–normal 1.84 s; foreign–normal 1.84 s; familiar–unusual 1.75 s; foreign–unusual 1.79 s.

3.2. fMRI results

3.2.1. Localiser experiment

The right temporo-parietal junction was identified in 12/12 subjects (average peak voxel [54 −54 24]), the left temporo-parietal junction in 8/12 subjects (average peak voxel [−48 −69 21]), the medial pre-frontal cortex in 11/12 subjects (average peak voxel [0 60 12]), and the posterior cingulate in 11/12 subjects (average peak voxel [3 60 24]). Sample regions of interest are shown in Fig. 1.

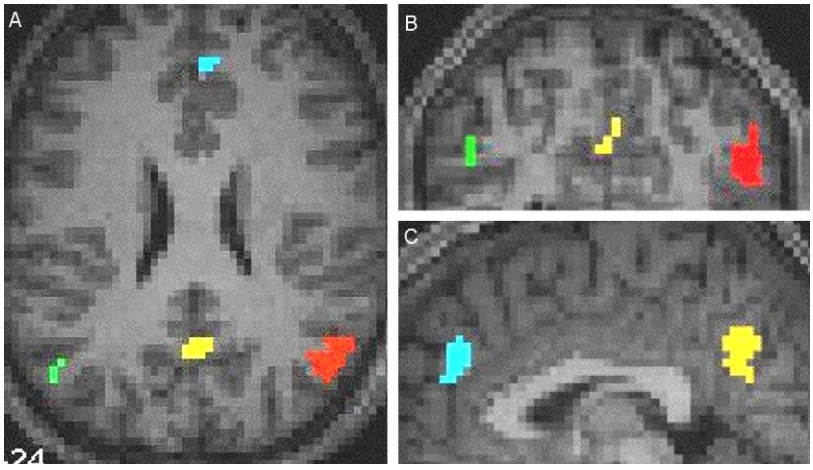

Fig. 1. Four 'Theory of Mind' regions of interest (ROIs) in a single representative subject. ROIs were defined as contiguous voxels in which the response was higher when subjects read stories about beliefs than when subjects read logically similar stories about photographs ($p < 0.0001$, uncorrected). Red = right temporo-parietal junction (RTPJ). Green = left TPJ. Cyan = medial prefrontal cortex (MPFC). Yellow = posterior cingulate (PC). (A) Axial slice, $z = 24$. (B) Coronal slice, $y = -60$. (C) Saggital slice, $x = 4$ (midline).

3.2.2. Background

In the first segment of each story, subjects read a description of a character's background that was either 'Foreign' or 'Familiar'. We compared the average PSC when only the background information was on the screen (Fig. 2). There was no effect of the background manipulation in either the RTPJ (familiar background PSC: 0.22, foreign PSC: 0.27, $t(1, 11) = 0.57$, $p > 0.5$) or the PC (familiar PSC: 0.58, foreign PSC: 0.63, $t(1, 10)=.42$, $p > 0.5$). The LTPJ did respond significantly more to Foreign than to Familiar backgrounds (familiar PSC: 0.62, foreign PSC: 0.86, $t(1, 7) = 2.85$ $p < 0.03$), and there was a trend in the same direction in the MPFC (familiar PSC: -0.02, foreign PSC: 0.21, $t(1, 10) = 1.75$, $p = 0.1$).

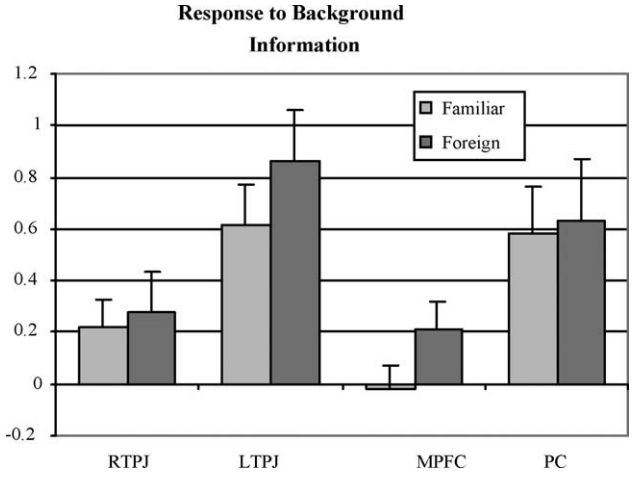

Fig. 2. Percent signal change in four 'Theory of Mind' regions of interest, while subjects read about the social background of the protagonist. Only the left TPJ showed a significantly higher response to 'Foreign' than 'Familiar' backgrounds ($p < 0.03$), although the medial prefrontal cortex response showed a trend in the same direction.

To measure the overall response of each region to a protagonist's background (social information with no mental state content), we compared the response of each ROI during the first six seconds of the stimulus when mental state information was delayed (the current experiment) and mental state information was available immediately (the belief stories from the localiser experiment).

The effect of delay was highly significant in the right TPJ ($t(1,11)=9.48$, $p < 0.001$, paired-samples t-test, Fig. 3a) and in the MPFC ($t(1,10)=3.25$, $p < 0.01$). There was no effect of delay in the left TPJ ($t(1,7) = 1.71$, $p > 0.1$) or in the posterior cingulate ($t(1,10) = 0.9$, $p > 0.3$). Only the effect in the RTPJ survived a Bonferroni adjustment for multiple comparisons. Repeated measures ANOVAs revealed that the MPFC, LTPJ, and posterior cingulate all showed significantly less effect of delay than the RTPJ (interaction region by delay, all $F > 10.0$, all $p < 0.01$, Fig. 3b).

3.2.3. Mental State

In the second segment of each text, a description of what the character wanted or believed was added to the story. The character's mental state was either normal (with respect to our subjects) or norm-violating. Combined with the background information, this yielded a 2×2 design. We compared the average PSC for the 6 s when the character's mental state became available for each region, using a 2×2 ANOVA (background by mental state).

The RTPJ response was higher when reading about the mental states of a person from a foreign background ($F(1, 11) = 8.91$ $p < 0.05$), but this main effect was mediated by a strong interaction with the mental state condition ($F(1, 11) = 18.71$ $p < 0.001$). That is, in the RTPJ, the BOLD response was higher to norm-violating mental states in characters from a familiar background, and to normal mental states in characters from a foreign background, than the reverse pairs (Fig. 4).

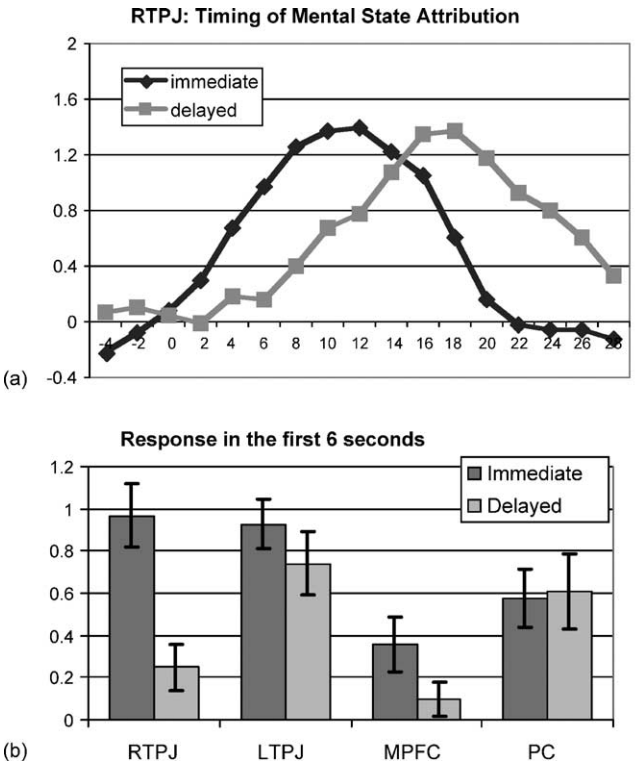

Fig. 3. (a) BOLD response in the RTPJ, relative to fixation (measured in percent signal change (PSC)), when mental state information was available from the time of onset of the stories ('immediate'), and when mental state information was delayed 6 s ('delay') while subjects read about the protagonist's background, averaged across both 'Familiar' and 'Foreign' backgrounds. The RTPJ response was strongly selective for mental state information. Time is shown on the *x*-axis. (b) Percent signal change in four 'Theory of Mind' regions of interest, during the first 6 s of stories about people (corresponding to timepoints 4, 6, and 8, above, to allow for the hemodynamic lag). Dark bars ('immediate') show the response while subjects read about a protagonist's mental states. Light bars ('delayed' mental states) show the response while subjects read about the protagonist's background.

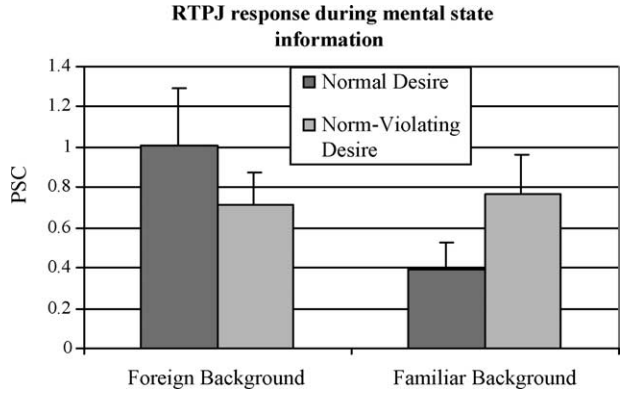

Fig. 4. Response of the RTPJ when the protagonist's mental state was described. For each background ('Familiar' versus 'Foreign') the RTPJ's response is enhanced to the incongruent mental states ('Norm-Violating', and 'Normal', respectively; interaction, $p < 0.001$).

There was a similar interaction of background and mental state in the posterior cingulate ($F(1, 10) = 7.57$, $p < 0.02$). But in the LTPJ, the same ANOVA revealed only a significant main effect of background (foreign greater than familiar, $F(1, 7) = 16.36$, $p < 0.01$), that did not interact significantly with the character's mental state. In the MPFC there were no significant main effects or interactions in the response during this time period. Only the interaction of background and mental state in the RTPJ survived a Bonferroni adjustment for multiple comparisons.

4. Discussion

Our subjects appeared to be very adept – even formulaic – at applying the ToM maxim that 'people's feelings have to be predicted from their own subjective desires' (Terwogt & Rieffe, 2003). A previous behavioural study found that when asked to 'really consider' the protagonist's feelings, subjects tended to overrule the protagonist's stated desire under very specific circumstances: when a protagonist from a 'Familiar' background professed a 'Norm-violating' desire. Thus, the subjects were likely to say that Andrew, their friend from high school, would really be hurt if his wife had an affair, even though he had said that he wanted her to do so (Terwogt & Rieffe, 2003, see Appendix A). By contrast, subjects in the scanner were equally fast and accurate when predicting the feelings of protagonists from all four groups (2 background × 2 mental state).

Nevertheless, the neural data suggest that our subjects were attempting to form an integrated impression of the protagonist in each story, and to resolve inconsistencies between expectations based on the protagonist's social background and her stated belief or desire. One brain region – the RTPJ – fulfilled each of the predictions for the neural substrate of Theory of Mind: (1) the BOLD response in the RTPJ was low while subjects read descriptions of a protagonist's social background, and increased only once the mental state of the protagonist was described, (2) the low response to background information was not modulated by the familiarity of the described background, and (3) once mental state information was available, the BOLD response in the RTPJ was enhanced when the protagonist's background and mental state were incongruent (e.g. a protagonist from a 'Foreign' background who professed a 'Normal' mental state) relative to when the background and mental state were congruent.

None of the other brain regions investigated here (the LTPJ, PC, and MPFC) showed as clear and unambiguous a response profile. In particular, the medial prefrontal cortex, which some authors have proposed as the unique site of true ToM reasoning (e.g. Gallagher & Frith, 2003), did not clearly fulfil any of the three predictions. (1) The response of the MPFC was lower in response to a protagonist's background than to the protagonist's mental state, but this difference did not survive a correction for multiple comparisons,

and was significantly smaller than the analogous effect in the RTPJ. (2) There was a trend in the MPFC towards a higher BOLD response to 'Foreign' than to 'Familiar' backgrounds. (3) There was no significant effect (or interaction) of the story condition during the 'Mental State' section of the stories.

Taken together, these results refute the suggestion that the MPFC is the unique neural substrate of Theory of Mind while the RTPJ serves only a precursor function such as the detection of agents or the processing of any socially relevant stimulus (e.g. Gallagher & Frith, 2003). Rather, the response of the RTPJ is highly specific to the attribution of mental states, while the MPFC may be less so. Recent lesion studies are consistent with this conclusion. Patients with selective damage to medial prefrontal cortex was found to be unimpaired on tests of Theory of Mind (Bach, Happe, Fleminger, & Powell, 2000; Bird, Castelli, Malik, Frith, & Husain, 2004), and three patients with damage to the left temporo-parietal junction were found to be selectively impaired in Theory of Mind (Samson, Apperly, Chiavarino, & Humphreys, 2004, although note that the tests used to assess the two patient populations differ considerably). Patients with selective damage to right temporo-parietal junction have not yet been tested, to our knowledge.

Patients with RTPJ damage may be particularly informative because while both left and right temporo-parietal junctions are consistently implicated in Theory of Mind, the current study suggests that there may be a laterality effect in the selectivity of their respective contributions. Unlike the RTPJ, the left TPJ showed a robust response to social background information that was not significantly different from the response during mental state attribution itself. Also, the left TPJ response to Foreign backgrounds was significantly higher than to Familiar backgrounds. We speculate that the left TPJ plays a broader role in the attribution of (enduring) socially relevant traits, while the RTPJ is restricted to the attribution of relatively transitive mental states. Consistent with this conclusion, following damage to left temporo-parietal junction (or nearby posterior superior temporal sulcus), patients were selectively impaired in the attribution of personality traits (but not emotional states) to point-light walkers (Heberlein, Adolphs, Tranel, & Damasio, 2005).

In addition to being strongly selective for the attribution of mental states, the response of the RTPJ was enhanced when the protagonist's background and mental state were incongruent. These results are incompatible with one simple interpretation of Simulation Theory (ST) according to which other minds are represented fundamentally in terms of their similarity to the perceiver's own mind. The response of the RTPJ was not linearly related to the similarity between the minds of the observer and the protagonist. Instead, the RTPJ appeared to reflect a process of constructing a coherent model of the protagonist's mind, without reference to the subject's own mental states. In fact, no region of the brain showed the ST-predicted linear relationship (that is, main effects of the observer-protagonist similarity in both the background and the mental state, going in the same direction).

Of course, other versions of the Simulation Theory remain unscathed. For instance, a simulation theorist could argue that incongruent background–mental state pairs are more difficult to simulate than congruent pairs, or that the response of the RTPJ reflects second-order similarity between the subject and the protagonist, at the level of internal coherence. Therefore, these results do not rule out a Simulation Theory model of reasoning about other minds, but only constrain its possible neural instantiation.

Interestingly, the effect of incongruence was apparent in the RTPJ, a putative domain specific substrate of Theory of Mind. Previous behavioural research has suggested that inconsistency resolution (encoding counter-stereotypical traits – e.g. an elderly person who is daring) depends on domain general executive function, and is disrupted by standard executive tasks like random number generation (Macrae, Bodenhausen, Schloerscheidt, & Milne, 1999). Our current design did not allow us to test the extent of the effect of incongruence across the whole brain, but these results do suggest that future neuroimaging studies could help to illuminate the process of building a coherent model of another mind. The RTPJ was the most selective of the brain regions investigated here, but must undeniably form only one component of the neural substrate of reasoning about other minds. A critical topic for future research will be to characterise the distinct and interacting contributions of other domain specific brain regions (possibly including the three investigated here) as well as of brain regions involved in domain general functions such as inhibitory control and executive function, in solving complex realistic Theory of Mind tasks (Saxe, 2005; Samson, Apperly, Kathirgamanathan, & Humphreys, 2005).

The current results also highlight the importance of broadening the scope of research on lay psychology (see also Nichols & Stich, 2003). Most work on Theory of Mind has investigated attributions of isolated transient mental states – e.g. beliefs, desires, and emotions. A separate tradition within social psychology has been concerned with the attribution of coherent, enduring, dispositional properties of a person, like personality traits (Gilbert, 1998; Malle, 1999), and the processes by which perceivers attempt to form an integrated, coherent perception of the target individual's personality (e.g. Asch, 1946). Future work should seek to integrate the insights from these two traditions into a single richer framework for understanding how human beings make sense of one another.

Acknowledgements

This work was funded by grant NIMH 66696. Thanks to Andrew Baron, Daniel Gilbert, Susan Carey, Lindsey Powell, Yuhong Jiang, Laura Schulz, Tania Tzelnic, Jason Mitchell, and especially to Nancy Kanwisher.

Appendix A

(1)

Familiar background

Your friend Lisa lives with her parents in New York City. She has a good job with a good salary and she is going to rent an apartment downtown.

Normal desire

Lisa is really looking forward to living alone. Yesterday she went to see a new place and fell totally in love with it. She desperately wants to live there.

1st outcome

The real estate agent called Lisa today to tell her that the apartment she looked at is available. Her parents say she should take it.

Foreign background

Your friend Lisa lives with her parents in rural Western Ireland. In that traditional community, it is uncommon and even suspicious for a woman to live alone.

Norm-violating desire

Lisa is happy living with her parents, and does not want the independence and responsibility of her own place. She would rather let her parents make the rules.

2nd outcome

The real estate agent called Lisa today to tell her that there are no apartments available right now.

(2)

Familiar background

Your friend Andrew, from high school, lives in Philadelphia. He and his wife have always had an excellent relationship. They almost never fight.

Normal desire

Andrew once confided in you that he really hates the idea that his wife might ever have an affair. Monogamy is very important to him.

1st outcome

Andrew is visiting you for dinner one evening and tells you that he asked his wife, and she said she will never sleep with another man.

Foreign background

Your friend Andrew, from high school, and his wife have become involved with a cult. Within their cult, extramarital relationships are accepted and occur often.

Norm-violating desire

Andrew once confided in you that he would find it fun if his wife, outside of their marriage, started a relationship with another man.

2nd outcome

Andrew is visiting you for dinner one evening and tells you that he just found out that his wife has been sleeping with another man.

References

Asch, S. E. (1946). Forming Impressions of Personality. *Journal of Abnormal and Social Psychology*, *41*, 258–290.

Bach, L. J., Happe, F., Fleminger, S., & Powell, J. (2000). Theory of mind: independence of executive function and the role of the frontal cortex in acquired brain injury. *Cognitive Neurospsychiatry*, *5*(3), 175–192.

Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: the case of information overload. *Journal of Personality and Social Psychology*, *49*, 1129–1146.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. Oxford: Oxford University Press.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2003). FMRI responses to video and point-light displays of moving humans and manipulable objects. *Journal of Cognitive Neuroscience*, *15*(7), 991–1001.

Bem, D. J. (1967). Self-perception: an alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–2000.

Bird, C. M., Castelli, F., Malik, O., Frith, U., & Husain, M. (2004). The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. *Brain*, *127*(Pt 4), 914–928.

Brunet, E., Sarfati, Y., et al. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *Neuroimage*, *11*, 157–166.

Castelli, F., Happe, F., Frith, U., & Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage*, *12*(3), 314–325.

Dennett, D. (1978). Beliefs about beliefs. *Behavioural and Brain Sciences*, *1*, 568–570.

Downing, P. E., Jiang, Y., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473.

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., et al. (1995). Other minds in the brain: a functional imaging study of 'theory of mind' in story comprehension. *Cognition*, *57*, 109–128.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*, 459–473.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends Cognitive Science*, *7*, 77–83.

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.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed.). New York: McGraw Hill.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *Neuroreport*, *6*(13), 1741–1746.

Goldman. (1992). In defense of the simulation theory. *Mind and Language*, *7*(1–2), 104–119.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. MIT Press, 288 pp.

Gordon, R. (1998). Radical simulation. In *In theories of theories of mind*. Carruthers & Smith.

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic with-in category identification. *Nature Neuroscience*, *7*(5), 555–562.

Grossman, E., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, *35*, 1167–1175.

Hamilton, D. L. (1988). Causal attribution from an information-processing perspective. In D. Bar-Tal & A. W. Kruglanski (Eds.), *The social psychology of knowledge* (pp. 359–385). New York: Cambridge University Press.

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review*, *103*, 336–355.

Happe, F. (2003). Theory of mind and self. *Annals of the New York Academy of Sciences*, *1001*, 134–144.

Harris, P. (1992). From simulation to folk psychology: the case for development. *Mind and Language*, *7*, 120–144.

Heal, J. (1998). Simulation, theory and content. In *In theories of theories of mind*. Carruthers & Smith.

Heberlein, A. S., Adolphs, R., Tranel, D., & Damasio, H. (2005). Cortical regions for judgements of emotions and personality traits from point-light walkers. *Journal of Cognitive Neuroscience*, *16*(7).

Higgins, E. T., & Bargh, J. A. (1987). Social perception and social cognition. *Annual Review of Psychology*, *38*, 369–425.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, *3*(1), 80–84.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

LaBar, K. S., Crupain, M. J., Voyvodic, J. T., & McCarthy, G. (2003). Dynamic perception of facial affect and identity in the human brain. *Cerebral Cortex*, *13*(10), 1023–1033.

Leslie, A., & Thaiss, L. (1992). Domain specificity in conceptual development. *Cognition*, *43*, 225–251.

Macrae, C. N., Bodenhausen, G. V., Schloerscheidt, A. M., & Milne, A. B. (1999). Tales of the unexpected: executive function and person perception. *Journal of Personality and Social Psychology*, *76*(2), 200–213.

Malle, B. (1999). How people explain behaviour: a new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23–48.

Nichols, S., & Stich, S. (2003). *Mindreading*. Oxford: Clarendon Press.

Nichols, S., Stich, S., Leslie, A., & Klein, D. (1995). Varieties of off-line simulation. In P. Carruthers & P. Smith (Eds.), *Theories of theories of mind* (pp. 39–74). Cambridge: Cambridge University Press.

Pelphrey, K., Singerman, J., Allison, T., & McCarthy, G. (2003). Brain activation evoked by perception of gaze shifts: the influence of context. *Neuropsychologia*, *41*, 156–170.

Pessoa, L., & Ungerleider, L. G. (2004). Neural correlates of change detection and change blindness in a working memory task. *Cerebral Cortex*, *14*(5), 511–520.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12–21.

Ruby, P., & Decety, J. (2003). Effect of subjective perspective taking during simulation of action: a PET investigation of agency. *Nature Neuroscience*, *4*(5), 546–550.

Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nature Neuroscience*, *7*(5), 499–500.

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: a case of a selective deficit in inhibiting self-perspective. *Brain*, *128*, 1102–1111.

Saxe, R. (2005). Four brain regions for one Theory of Mind? In J. Cacioppo (Ed.), *People thinking about people*. MIT Press.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87–124.

Saxe, R., Jamal, N., & Powell, L. (2005). My body or yours? The effect of visual perspective on cortical body representations. *Cereb Cortex* [e-published].

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: fMRI investigations of theory of mind. *Neuroimage*, *9*(4), 1835–1842.

Saxe, R., Xiao, D. K., Kovacs, G., & Perret, D. I. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, *42*(11), 1435–1446.

Scholl, B., & Leslie, A. (2001). Minds, modules, and meta-analysis. *Child Development*, *72*(3), 696–701.

Schultz, R. T., Grelotti, D. J., Klin, A., Kleinman, J., Van der Gaag, C., Marois, R., & Skudlarski, P. (2003). The role of the fusiform face area in social cognition: implications for the pathobiology of autism. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1430), 415–427.

Stich, S., & Nichols, S. (1992). Folk psychology: simulation or tacit theory. *Mind and Language*, *7*(1), 35–71.

Terwogt, M. M., & Rieffe, C. (2003). Stereotyped beliefs about desirability: implications for characterizing the child's theory of mind. *New Ideas in Psychology*, *21*(1), 69–84.

Vaina, L. M., Gryzwacz, N. M., Saiviroonporn, P., LeMay, M., Bienfang, D. C., & Cowey, A. (2001). Can spatial and temporal motion integration compensate for deficits in local motion mechanisms? *Neuropsychologia*, *41*(13), 1817–1836.

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.

Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., Wright, C. I., & Rauch, S. L. (2001). A functional MRI study of human amygdala responses to facial expressions of fear versus anger. *Emotion*, *1*(1), 70–83.

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron*, *30/40*(3), 655–664.

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.

Winston, J. S., O'Doherty, J., & Dolan, R. J. (2003). Common and distinct neural responses during direct and incidental processing of multiple facial emotions. *Neuroimage*, *20*(1), 84–97.

Wyer, R. S., & Gordon, S. E. (1982). The recall of information about persons and groups. *Journal of Experimental Social Psychology*, *20*, 445–469.