Neuropsychologia 48 (2010) 2658–2664

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind

Liane Young*, David Dodell-Feder, Rebecca Saxe

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, USA

A R T I C L E I N F O

Article history:
Received 4 November 2009
Received in revised form 30 April 2010
Accepted 6 May 2010
Available online 12 May 2010

Keywords:
Social cognition
Theory of mind
Attention
Functional magnetic resonance imaging
Temporo-parietal junction

A B S T R A C T

Functional magnetic resonance imaging (fMRI) studies have demonstrated a critical role for a cortical region in the right temporo-parietal junction (RTPJ) in "theory of mind" (ToM), or mental state reasoning. In other research, the RTPJ has been implicated in the deployment of attention to an unexpected stimulus. One hypothesis ("attention hypothesis") is that patterns of RTPJ activation in ToM tasks can be fully explained by appeal to attention: stimuli that apparently manipulate aspects of ToM are in fact manipulating aspects of attention. On an alternative hypothesis ("ToM hypothesis"), functional regions identified by ToM tasks are selective for ToM, and not just for any unexpected stimulus. Here, we used fMRI to test these competing hypotheses: are brain regions implicated in ToM, including the RTPJ, LTPJ, and precuneus, recruited specifically for mental states, or for any unexpected stimulus? We first identified brain regions implicated in ToM, using a standard paradigm: participants read stories about false beliefs and false physical representations (e.g., outdated photographs). Participants also read a new set of stories describing mental or physical states, which were unexpected or expected. Regions of interest analyses revealed a higher response in the RTPJ, LTPJ, and precuneus, for mental versus physical stories, but no difference for unexpected and expected stories. Whole-brain random effects analyses also revealed higher activation in these regions for mental versus physical stories. This pattern provides evidence for the ToM hypothesis: the response in these functional regions is selective for mental state content, whether that content is unexpected or expected.

1. Introduction

Functional magnetic resonance imaging (fMRI) studies have demonstrated a role for a cortical region in the temporo-parietal junction (TPJ) in "theory of mind" (ToM), the ability to represent and reason about mental states, such as thoughts and beliefs (Fletcher et al., 1995; Gallagher et al., 2000; Samson, Apperly, Chiavarino, & Humphreys, 2004; Saxe & Powell, 2006). For example, the blood oxygen level dependent (BOLD) response in the right TPJ (RTPJ), left TPJ (LTPJ), and precuneus (PC) is significantly higher when participants read stories explicitly describing or requiring inferences about mental states such as false beliefs as compared to when participants read stories about physical states such as false or outdated signs, maps, or photographs (Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007; Perner, Aichhorn, Kronbichler, Wolfgang, & Laddurner, 2006; Saxe & Kanwisher, 2003).

A separate body of work has suggested a role for a region in the TPJ, especially in the right hemisphere, in exogenously cued attention, or the reorienting of attention to an unexpected stimulus. fMRI studies show that the response in this region is significantly higher during the detection of low-frequency targets (Bledowski, Prvulovic, Goebel, Zanella, & Linden, 2004; Downar, Crawley, Mikulis, & Davis, 2000) or targets that appear in unexpected locations as in "invalidly cued" trials (Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000; Vossel, Weidner, Thiel, & Fink, 2009) of a Posner cueing paradigm (Posner, Walker, Friedrich, & Rafal, 1984). These results suggest that this cortical region in the right TPJ functions at least in part to deploy attention to unexpected or surprising stimuli. Furthermore, damage to this region leads to a deficit in reorienting of attention (Friedrich, Egly, Rafal, & Beck, 1998) and to left hemifield spatial neglect (Vallar, Bottini, Rusconi, & Sterzi, 1993; Vallar & Perani, 1986).

How should these two lines of research be integrated? Two sets of competing hypotheses have emerged. The first hypothesis ("attention hypothesis") is that patterns of TPJ activation, especially RTPJ activation, found in ToM tasks can be explained away by appeal to attention: stimuli that apparently manipulate aspects of ToM are in fact manipulating aspects of attention. In other words, stimuli designed to require more ToM may have elicited enhanced

* Corresponding author at: Department of Brain and Cognitive Sciences, MIT, 43 Vassar Street, Building 46, Room 4021, Cambridge, MA 02139, USA.
Tel.: +1 617 324 2890; fax: +1 617 324 2890.
E-mail address: lyoung@mit.edu (L. Young).

0028-3932/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2010.05.012

RTPJ activation only because these stimuli are also unexpected, and require integrating inconsistent information in "elaborate inference processes" (Ferstl, Neumann, Bogler, & von Cramon, 2008; Virtue, Parrish, & Jung-Beeman, 2008). Standard false belief tasks, for example, require participants to switch attention multiple times between at least two locations. The Sally–Anne task depicts the following situation: (1) Sally places her ball in a basket (location 1), and then leaves the room, (2) Anne enters the room, and moves her ball to the box (location 2), (3) Sally returns to retrieve her ball. Participants are asked to predict where Sally will look for her ball. Thus, the false belief task may require participants to attend to the unexpected switch in the object's location, to reorient attention between the two locations, and to integrate the inconsistent locations of the ball over time. In the standard control task, participants make judgments about physical representations that have become false or outdated such as "false photographs". The control events also involve an unexpected transfer of an object between two locations, and require reorienting attention between two locations. However, it is not easy to ascertain whether the "unexpectedness" of the two kinds of events is truly matched; it remains possible that the false belief stories engage, and therefore reorient, attention differently or more effectively than the control stories do.

As evidence for the attention hypothesis, Buccino et al. (2007) point out that a region near the TPJ (in the posterior superior temporal suclus, pSTS) shows higher metabolic activity when people observe unexpected or inconsistent human actions, relative to expected or consistent human actions (Grezes, Frith, & Passingham, 2004; Pelphrey, Morris, & McCarthy, 2004). They write that "although both the explanations for the activation of the temporo-parietal regions, the one based on theory of mind and that one based on attention, may be valid, we are inclined to prefer the attentional explanation because the major feature of the non-intended actions used in [those experiments] was their unexpectedness" (Buccino et al., 2007).

An alternative hypothesis ("ToM hypothesis") is that patterns of TPJ activation, especially RTPJ activation, found in ToM tasks cannot be explained by appeal to attention: instead, functional regions identified by ToM tasks (e.g., false belief versus false photograph) are selective for ToM, and not simply any unexpected stimulus requiring more attention. The ToM hypothesis is supported by recent fMRI work showing anatomically close but distinct cortical regions of the RTPJ that support distinct cognitive functions; that is, the region of the RTPJ that is recruited in ToM tasks is selective for ToM, while a nearby but distinctive region is involved in the reorienting of attention to unexpected stimuli (Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009).

One approach to resolving these two hypotheses ("attention hypothesis" and "ToM hypothesis") is to ask whether the brain regions recruited by low-level attentional reorientation and high-level ToM actually occupy the same region of cortex, near the right temporo-parietal junction. The first study to test this question reported anatomical overlap between the regions of the RTPJ that support ToM (i.e. in a false belief task) and low-level exogenous attention (i.e. in a Posner cueing task) in the same individuals (Mitchell, 2008). However, a subsequent study, using higher resolution imaging a bootstrap analysis, found a small but reliable separation between the peaks of these two functional regions in higher resolution images (Scholz et al., 2009), consistent with evidence from a recent meta-analysis (Decety & Lamm, 2007).

An alternative approach to the two hypotheses ("attention hypothesis" and "ToM hypothesis") is to test directly whether the activation patterns observed during ToM tasks can be explained away by differences in *high-level attention*. The current study takes this second approach, using high-level verbal stimuli in a single paradigm. Notably, prior research described above has focused on the relationship between ToM and exogenous (i.e. stimulus-driven) visual attention, relying on low-level sensory stimuli that are unexpected in virtue of their frequency or spatial location (Corbetta et al., 2000). If false belief stimuli recruit the RTPJ in virtue of their "unexpectedness", though, the expectations that are elicited, and violated, must be of a higher level and more abstract kind of expectation about the events described in the verbal vignettes (Ferstl et al., 2008). In the current study we therefore investigate the relationship between ToM and higher level attention. We manipulate the expectedness (validated with subjective measures) of high-level verbal stimuli describing both mental states and physical states. Both mental and physical states were unexpected (or expected) with respect to common knowledge. For example, unexpected mental stories featured protagonists with unlikely desires (e.g., the desire to make pesto sauce with chocolate and marijuana) or patently false beliefs (e.g., that watering the house plants will make them burst into flames). Correspondingly, physical states were designed to be unexpected with respect to the average participant's knowledge of the real world (e.g., the water from a tap tastes like milk chocolate).

We first identified brain regions implicated in ToM, including the RTPJ, LTPJ, precuneus (PC), and medial prefrontal cortex (MPFC), using a standard paradigm: participants read stories about false beliefs and outdated physical representations (e.g., false photographs). Participants then read a new set of stories describing mental or physical states, which were unexpected or expected. On the attention hypothesis, regions recruited for a false belief task should differentiate between unexpected and expected events, in general. On the ToM hypothesis, these regions should differentiate only between mental and physical stories, and not between unexpected and expected stories.

2. Methods

2.1. Participants

Seventeen naïve right-handed adults (aged 18–31, 7 females) participated in the study for payment. All participants were native English speakers, had normal or corrected-to-normal vision, and gave written informed consent in accordance with the requirements of the internal review board at MIT.

2.2. Stimuli

Stimuli consisted of two sets of 96 stories (Supplementary material): (1) stories describing mental states that were either expected or unexpected and (2) stories describing physical events, objects, or states that were either expected or unexpected (Fig. 1). Word count was matched across conditions (mean ± SD for the mental condition: 13 ± 2; physical condition: 13 ± 3; unexpected condition: 13 ± 2; expected condition: 13 ± 3) such there was no significant difference in word count between mental and physical stories ($F(1,188) = 2.83$, $p = 0.09$, partial $h^2 = 0.02$) or between unexpected and expected stories ($F(1,188) = 0.30$, $p = 0.86$, partial $h^2 < 0.001$). A question accompanied each version (expected and unexpected) of the mental and physical stories. Word count for the mental and physical questions also did not differ significantly (mean ± SD for the mental questions: 10 ± 2; physical questions: 10 ± 2; $F(1,94) = 0.009$, $p = 0.92$, partial $h^2 < 0.001$). The expectedness or unexpectedness of the stories was validated with a post-scan questionnaire in which participants viewed the stories presented during the scan and rated them on a 7-point scale (1 = not at all surprising; 7 = very surprising).

In the scanner, stories were presented for 6 s, followed by a question for 6 s and finally 10 s of fixation on a black screen. During the question portion of the trial, participants judged how likely it would be for the story protagonist to hold another specific belief or desire for the mental stories, and how likely it would be for the physical state or object in the story to have another specific property for the physical stories, using four buttons: 1 = very unlikely, 4 = very likely (Fig. 1). Due to technical error, behavioral data were not collected for one participant.

Participants saw either the expected or the unexpected version of the mental and physical stories for a total of 48 stories. Stories were presented in a pseudo-randomized order with the order of conditions counterbalanced across runs and participants. Twelve stories (three stories per condition) were presented during each of four runs for a total time of 18 min and 8 s. The text of each story was presented in a white 36-point font on a black background via Matlab 7.6 running on an Apple MacBook Pro. The scan session also included four runs of a ToM functional localizer, contrasting stories about mental states (e.g., false beliefs) and stories about physical states (e.g., false photographs; see Saxe & Kanwisher, 2003, Experiment 2).