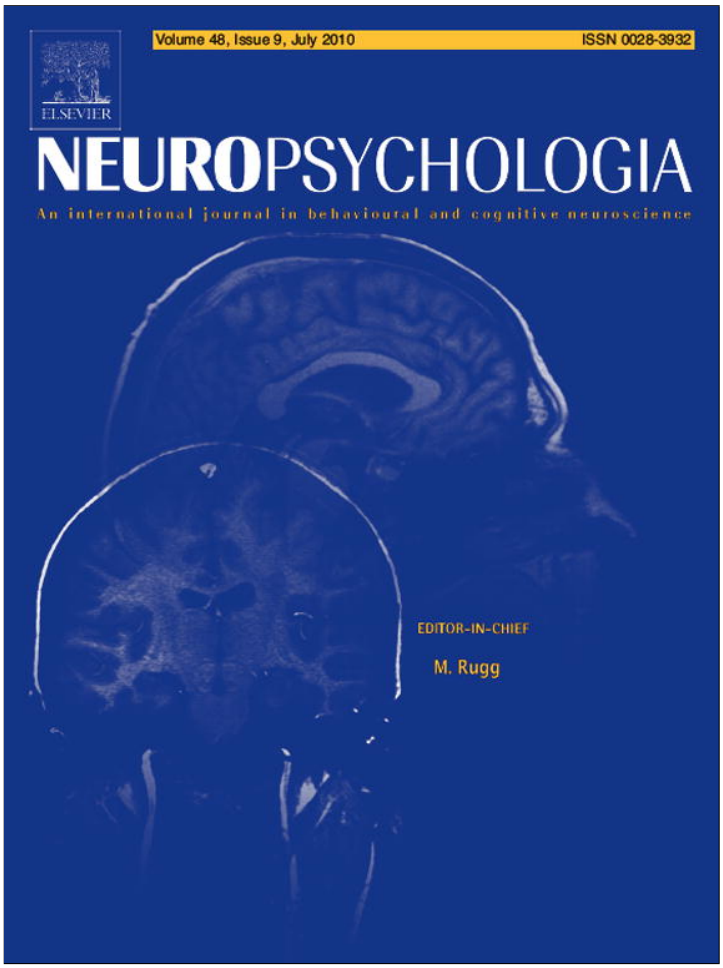

Neuropsychologia 48 (2010) 2658–2664

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

ARTICLE INFO

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

ABSTRACT

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ToM, using a standard paradigm: participants read stories about false beliefs and false physical representations (e.g., outdated photographs).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).

	Mental	Physical
Expected	Maya thinks the house plants will flower a few times if watered regularly.	The house plants will flower a few times a year if watered regularly.
Unexpected	Maya thinks the house plants will burst into flames if watered regularly.	The house plants will spontaneously burst into flames if watered regularly.

How likely is it that...

... Maya thinks the plants also need sunlight?	... the plants also need exposure to sunlight?

not at all 1 -- 2 -- 3 -- 4 very

Fig. 1. Stimuli design. Stimuli consisted of two sets of 48 stories: (1) "mental stories" describing mental states that were either expected or unexpected and (2) "physical stories" describing physical events, objects, or states that were either expected or unexpected. Participants judged how likely it would be for the protagonist to hold another specific mental state 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 (1 = very unlikely, 4 = very likely).

2.3. Imaging procedure

Participants were scanned at 3 T (at the MIT scanning facility in Cambridge, MA) using thirty 4-mm-thick near axial slices covering the whole-brain. Standard echoplanar imaging procedures were used (TR = 2 s, TE = 40 ms, flip angle = 90°).

FMRI data were analyzed using SPM2 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. Each participant's data were motion corrected and normalized onto a common brain space (Montreal Neurological Institute, MNI, template). Data were smoothed using a Gaussian filter (full width half maximum = 5 mm) and were high-pass filtered during analysis. The experiment used a block design and was modeled using a boxcar regressor.

Both whole-brain and tailored ROI analyses were conducted. Six ROIs were defined for each participant individually based on a whole-brain analysis of a localizer contrast and defined as contiguous voxels that were significantly more active ($p < 0.001$, uncorrected, $k > 10$) while the participant read the false belief stories, as compared with the false photograph stories: RTPJ, LTPJ, precuneus (PC), dorsal medial prefrontal cortex (DMPFC), middle MPFC (MMPFC), and ventral MPFC (VMPFC). All peak voxels are reported in MNI coordinates.

The responses of these ROIs were then measured while participants read the new stories from the current study. Within the ROI, the average percent signal change (PSC) relative to 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 and all blocks of the same condition). We then averaged across the time points during which the story and question was presented (4–14 s after story onset, to account for hemodynamic lag) to get a single PSC value for each region in each participant (Poldrack, 2006). This value was used in all analyses reported below.

3. Results

3.1. Behavioral results: post-scan questionnaire

Story expectedness ratings were analyzed using a 2 (mental versus physical) × 2 (unexpected versus expected) repeated measures ANOVA of participants' average ratings for each of the four conditions (Fig. 2). As expected, a significant difference in expectedness ratings was observed between the unexpected and expected stories ($F(1,16) = 689$, $p < 0.001$, partial $h^2 = 0.98$), but not between the mental and physical stories ($F(1,16) = 1.19$, $p = 0.29$, partial $h^2 = 0.07$), and there was no interaction. Paired samples t-tests revealed that unexpected stories were rated as more unexpected for both mental stories ($t(16) = 25.43$, $p < 0.001$) and physical stories ($t(16) = 20.71$, $p < 0.001$).

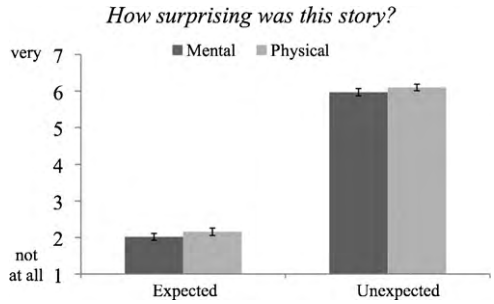

Fig. 2. Expectedness ratings. In a post-scan questionnaire, participants viewed the stories presented during the scan: mental stories (dark bars) and physical stories (light bars). Stories were rated on a 7-point scale (1 = not at all surprising; 7 = very surprising). Error bars represent standard error.

3.2. Behavioral results: scanner task

Likelihood ratings were analyzed using the same procedure as in the post-scan questionnaire. A difference in likelihood ratings was observed between the unexpected and expected stories ($F(1,15) = 10.34$, $p = 0.01$, partial $h^2 = 0.41$), but not between the mental and physical stories ($F(1,15) = 2.10$, $p = 0.11$, partial $h^2 = 0.16$), and there was no interaction. Paired samples t-tests revealed that questions for the expected stories were rated as more likely for both mental ($t(15) = 2.12$, $p = 0.04$) and physical stories ($t(15) = 3.34$, $p = 0.004$). No reaction time differences were observed (unexpected versus expected: $F(1,15) = 2.28$, $p = 0.15$, partial $h^2 = 0.13$; mental versus physical: $F(1,15) = 3.16$, $p = 0.10$, partial $h^2 = 0.17$).

3.3. fMRI results: functional localizer

A whole-brain random effects analysis of the ToM functional localizer data replicated results of studies using the same task (Saxe & Kanwisher, 2003), revealing a higher BOLD response during stories about false beliefs versus stories about false photographs, in the RTPJ, LTPJ, DMPFC, MMPFC, VMPFC, and PC ($p < 0.001$, uncorrected, $k > 10$). These ROIs were identified in individual participants at the same threshold: RTPJ (identified in 17 of 17 participants), LTPJ (16/17), PC (17/17), DMPFC (13/17), MMPFC (9/17), and VMPFC (11/17) (Table 1 and Fig. 3).

Table 1
Functional localizer experiment results.

Region	x	y	z
Individual ROIs			
RTPJ	54	−53	23
LTPJ	−52	−58	22
PC	2	−56	37
DMPFC	0	57	29
MMPFC	3	57	14
VMPFC	1	53	−12
Whole-brain contrast			
RTPJ	58	−52	28
LTPJ	−56	−52	26
PC	−2	−54	38
DMPFC	0	56	32
MMPFC	4	64	16
VMPFC	2	44	−20

Average peak voxels for ROIs in Montreal Neurological Institute coordinates. The "Individual ROIs" rows show the average peak voxels for individual participants' ROIs. The "Whole-brain contrast" rows show the peak voxel in the same regions in the whole-brain random effects group analysis.

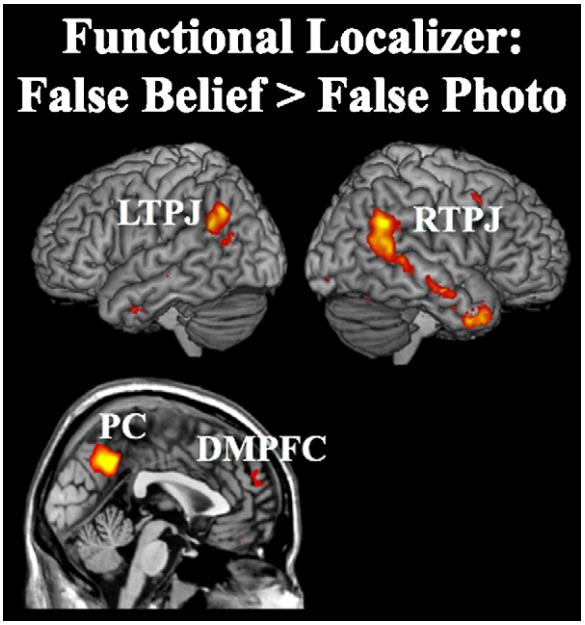

Fig. 3. Functional localizer results. Brain regions where the BOLD signal was higher for stories about mental representations (e.g., false beliefs) than stories about physical representations (e.g., false photographs; $N = 17$, random effects analysis, $p < 0.001$, uncorrected). These data were used to define regions of interest.

3.4. fMRI results: story task

The percent signal change (PSC) from rest in each of the ROIs was calculated for the time when the story and question were on the screen. The PSC in each ROI was then analyzed in a 2 (mental versus physical) × 2 (unexpected versus expected) repeated measures ANOVA.

3.4.1. TPJ and PC

A similar pattern of results was observed in the RTPJ, LTPJ and PC (Fig. 4). In these ROIs, the PSC was higher for mental versus physical content (RTPJ: $F(1,16) = 11.28$, $p = 0.004$, partial $h^2 = 0.41$; LTPJ: $F(1,15) = 30.00$, $p < 0.001$, partial $h^2 = 0.67$; PC: $F(1,16) = 10.71$, $p = 0.005$, partial $h^2 = 0.40$). The PSC in these regions did not discriminate between unexpected versus expected stories (RTPJ: $F(1,16) = 1.22$, $p = 0.29$, partial $h^2 = 0.07$; LTPJ: $F(1,15) = 2.61$, $p = 0.13$, partial $h^2 = 0.15$; PC: $F(1,16) = 0.06$, $p = 0.81$, partial $h^2 = 0.003$). There was no interaction. Paired samples t-tests revealed that the response in these regions discriminated between the mental and physical stories both when they were unexpected (RTPJ: $t(16) = 2.00$, $p = 0.06$; LTPJ: $t(15) = 5.57$, $p < 0.001$; PC: $t(16) = 2.63$, $p = 0.018$) and when they were unexpected (RTPJ: $t(16) = 3.00$, $p = 0.009$; LTPJ: $t(15) = 4.07$, $p = 0.001$; PC: $t(16) = 2.78$, $p = 0.013$). Critically, none of the ROIs discriminated between the unexpected and expected stories in the mental domain (RTPJ: $t(16) = 0.88$, $p = 0.39$; LTPJ: $t(15) = 1.63$, $p = 0.12$; PC: $t(16) = -0.67$, $p = 0.52$) or in the physical domain (RTPJ: $t(16) = 1.06$, $p = 0.31$; LTPJ: $t(15) = 1.17$, $p = 0.26$; PC: $t(16) = 0.35$, $p = 0.73$).

3.4.2. MPFC

In the DMPFC, the response was higher for mental versus physical content ($F(1,12) = 8.35$, $p = 0.014$, partial $h^2 = 0.41$) and unexpected versus expected stories ($F(1,12) = 7.92$, $p = 0.016$, partial $h^2 = 0.40$). There was no interaction. Paired samples t-tests revealed that the DMPFC response was higher for mental versus physical stories when they were expected ($t(12) = 2.90$, $p = 0.013$), but not when they were unexpected ($t(12) = 1.68$, $p = 0.12$). The DMPFC response did not discriminate between unexpected and expected in the mental domain ($t(12) = 1.29$, $p = 0.22$) or in the physical domain ($t(12) = 0.98$, $p = 0.35$). We observed no significant effects in the MMPFC and VMPFC in either analysis.

In sum, the response in the RTPJ, LTPJ, and PC showed a main effect of mental versus physical and no main effect of unexpected versus expected. Furthermore, these regions discriminated between mental and physical stories when they were expected and unexpected. By contrast, the response in the DMPFC showed both main effects of mental versus physical and unexpected versus expected, and also did not discriminate between mental versus physical stories when they were unexpected.

To explore the differences in the functional profiles of these ROIs, which produced significant effects, we conducted a 4 (region: RTPJ,

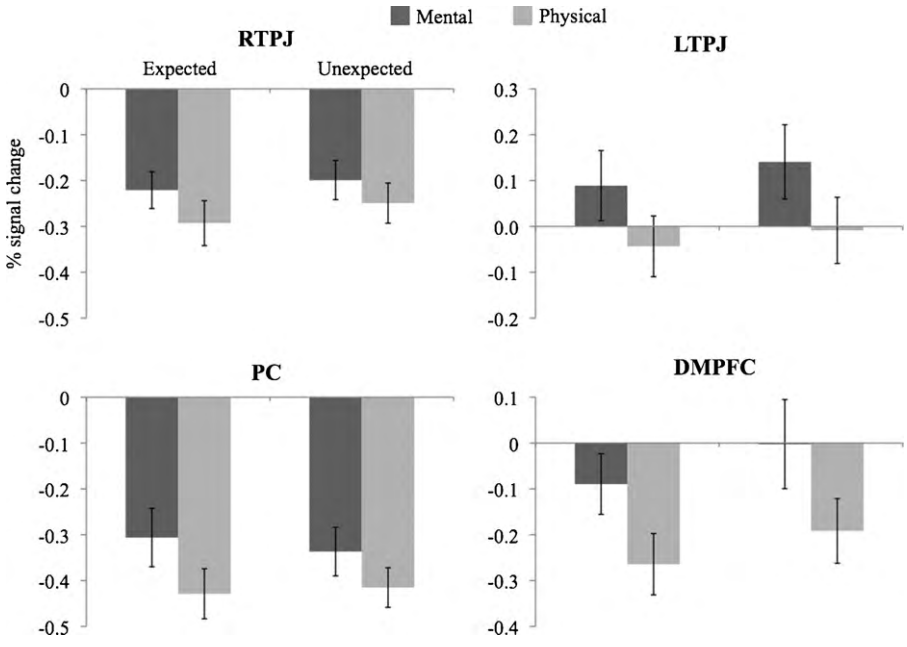

Fig. 4. Regions of interest analyses. Percent signal change (PSC) in the RTPJ, LTPJ, PC, and DMPFC when participants read unexpected and expected mental stories (dark bars) and physical stories (light bars). Error bars represent standard error.

Table 2
Story task whole-brain analysis results.

Region	x	y	z
Mental > physical			
RTPJ	62	−56	22
LTPJ	−54	−58	26
PC	2	−70	34
Unexpected > expected			
PC	−10	−70	32
ACC	−8	36	−6
VMPFC	0	44	−6

Average peak voxels for ROIs in Montreal Neurological Institute coordinates.

LTPJ, PC and DMPFC) × 2 (mental versus physical) × 2 (unexpected versus expected) ANOVA. This analysis revealed a main effect of mental versus physical ($F(1,12) = 12.36$, $p = 0.004$), a main effect of region ($F(3,36) = 7.31$, $p = 0.001$), and an interaction between region and unexpected versus expected ($F(3,36) = 3.55$, $p = 0.024$), suggesting differences in the functional profiles of the ROIs. We conducted additional region by function analyses for pairs of regions, in particular, to explore the differences observed above between the response in the DMPFC and the response in the RTPJ, LTPJ, and PC. A comparison of the RTPJ and DMPFC revealed a main effect of mental versus physical ($F(1,12) = 9.74$, $p = 0.009$) and an interaction between region and mental versus physical ($F(1,12) = 4.91$, $p = 0.047$). A comparison of the LTPJ and DMPFC revealed a main effect of mental versus physical ($F(1,12) = 12.83$, $p = 0.004$), and a main effect of unexpected versus expected ($F(1,12) = 6.08$, $p = 0.03$), but no interactions. A comparison of the PC and DMPFC revealed a main effect of region ($F(1,12) = 6.28$, $p = 0.03$), a main effect of mental versus physical ($F(1,12) = 8.3$, $p = 0.014$), an interaction between region and mental versus physical ($F(1,12) = 4.61$, $p = 0.05$), and an interaction between region and unexpected versus expected ($F(1,12) = 6.73$, $p = 0.02$).

Given our a priori hypotheses about the RTPJ, we also performed whole-brain random effects analyses for (1) the mental > physical contrast, and (2) the unexpected > expected contrasts ($p < 0.001$, uncorrected, $k > 10$; Table 2), and looked for clusters of activation within 20 mm of the peak coordinates of the RTPJ in the ToM functional localizer (peak [58 −52 28]). As predicted, we found a cluster in the RTPJ for the mental > physical contrast (peak [62 −56 22]), but no clusters for the unexpected > expected contrast. At a very lenient threshold ($p < 0.01$, uncorrected, $k > 5$), we found a small cluster of activity in the unexpected > expected contrast near the peak coordinates of the RTPJ in the ToM functional localizer [48 −46 22]. This weaker effect for the unexpected > expected contrast, compared to the mental > physical contrast, is consistent with prior work directly comparing in the same individuals theory of mind and attention to unexpected stimuli, as elicited in a Posner cueing task (Mitchell, 2008).

To further investigate these effects, we performed whole-brain conjunction analyses between the ToM localizer contrasts and (1) the mental > physical contrast, and then (2) the unexpected > expected contrast (Fig. 5). Each voxel counted as 'overlap' only if the contrast exceeded the *T*-threshold independently for both tasks ($T > 3.69$, $p < 0.001$). Consistent with the ROI analyses, the conjunction between the localizer and the mental > physical contrast revealed activation in the RTPJ, LTPJ and PC. The conjunction between the localizer and the unexpected > expected contrast revealed activation in the PC. At a very lenient threshold ($T > 2.58$, $p < 0.01$, uncorrected), the conjunction between the localizer and the unexpected > expected contrast revealed activation in the PC and a small cluster of activation in the RTPJ, consistent with the whole-brain random effects analysis, that did not overlap with the region of the RTPJ recruited by the mental > physical contrast (Fig. 5).

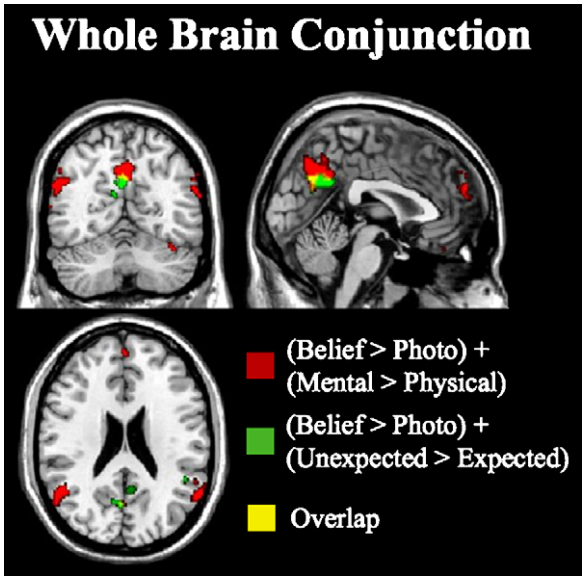

Fig. 5. Whole-brain conjunction analyses. Conjunction between the ToM localizer contrast (false belief > false photograph) and (1) the mental > physical contrast (shown in red) and (2) the unexpected > expected contrast (shown in green). Each voxel counted as 'overlap' only if the contrast exceeded the *T*-threshold independently for both tasks ($T > 2.58$, $p < 0.01$, uncorrected). The conjunction between the localizer and the mental > physical contrast revealed activation in the RTPJ, LTPJ, DMPFC and MMPFC. The conjunction between the localizer and the unexpected > expected contrast revealed activation in the PC and a small cluster of activation in the RTPJ, which did not overlap with the region of the RTPJ recruited by the mental > physical contrast. Overlap of these two independent conjunction analyses is shown in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

4. Discussion

Here we show, based on both regions of interest analyses and convergent whole-brain analyses, that brain regions for ToM are selectively recruited for mental versus non-mental content, and not simply for processing any unexpected stimulus. Brain regions recruited for a ToM task comparing false beliefs to false photographs (i.e. RTPJ, LTPJ, PC and DMPFC), were selectively recruited for a new set of stories about a range of mental states – both expected and unexpected. In support of the ToM hypothesis, but not the attention hypothesis, the RTPJ, LTPJ, and PC were sensitive only to the distinction between mental and physical states, and not to the distinction between unexpected and expected events, in either the mental or the physical domain. The DMPFC response was higher for both mental versus physical states and for unexpected versus expected events.

While prior research has focused on the relationship between ToM and attention to low-level sensory stimuli that were unexpected due to frequency or spatial location (Mitchell, 2008; Scholz et al., 2009), the current study manipulated attention to high-level verbal stimuli that were unexpected with respect to common knowledge (as verified by post-scan ratings). Manipulating "high-level attention" in the current study allowed us to probe ToM and an aspect of attention within a single paradigm, using the same high-level verbal stimuli. More specifically, we were able to test the hypothesis that if neural activation observed for false belief stories in ToM tasks can be explained away by enhanced attention to high-level verbal stimuli that are unexpected or surprising (Mitchell, 2008), then the same neural activation should be observed for the identical high-level verbal stimuli that are stripped of beliefs but that are equally unexpected or surprising. The current results do not support this attention hypothesis. Instead, these results show that these regions play a specific role in ToM. Future work, using

behavioral and neuroimaging methods, ought to explore the relationship between attention to unexpected sensory stimuli (as in Posner-type cueing paradigms) and attention to unexpected verbal stimuli (as in the current study).

Contrary to the current results, some previous research has suggested that activity in a region of the TPJ is modulated by manipulations of high-level attention. For example, an ERP study reports an N400 component localized to the posterior temporal neocortex near the temporo-parietal junction in response to semantically incongruous sentences (Simos, Basile, & Papanicolaou, 1997). A recent fMRI study has also suggested a role for the TPJ in the resolution of semantic incongruity and integrating inconsistent information: incoherent passages produced greater activation in the TPJ/STS (Ferstl et al., 2008). Interestingly, however, another fMRI study concluded the opposite: the right TPJ/STS is recruited "when comprehenders generate bridging inferences under highly predictable text conditions" (Virtue et al., 2008). Another line of fMRI research on action perception indicates activity in the right pSTS/TPJ for incorrect goal-directed actions versus correct goal-directed actions (Pelphrey et al., 2004), and incorrect versus correct predictions about the weight of an object (Grezes et al., 2004). Notably, however, across all the studies described above, activations patterns have been observed mostly in the STS. More recently, Buccino and colleagues reported increased RTPJ activity when participants viewed unintentional versus intentional actions (Buccino et al., 2007) and suggested that unintentional actions are unexpected, and therefore recruit the RTPJ. Another possibility, though, is that unintentional actions (or otherwise unexpected or incongruent actions, as in the studies above) provoke greater focus on the actor's mental states, including the actor's intentions. In sum, we suggest that the apparent inconsistency between these findings and the results of the currents study can be understood as follows: (1) the previously observed activations patterns have been mostly centered on the STS, and (2) the perception of unexpected or incongruent human actions may engage not only greater attention but also greater theory of mind, that is, reasoning about the beliefs and intentions of the actor.

An interesting aspect of the current results is that ToM brain regions were selective for mental state content but not for a particular feature of the mental state – whether it is unexpected or expected. Previous neuroimaging studies have suggested that the TPJ is recruited selectively for mental states over physical states (Perner et al., 2006; Saxe & Kanwisher, 2003), other socially relevant information such as physical traits or internal bodily sensations (Saxe & Powell, 2006), and the positive or negative effects of a protagonist's actions on another person, relevant for social or moral judgment (Young, Camprodan, Hauser, Pascual-Leone, & Saxe, 2010; Young & Saxe, 2008, 2009b). Nevertheless, the RTPJ response appears to be unaffected by manipulations of specific features of the mental states, including whether beliefs are true versus false (Jenkins & Mitchell, 2009; Young, Cushman, Hauser, & Saxe, 2007), justified or unjustified (Young, Nichols, & Saxe, in press), positive versus negative (Kliemann, Young, Scholz, & Saxe, 2008; Young et al., 2007; Young & Saxe, 2008, 2009a), and whether inferences about those states are "constrained" versus "open-ended" (Jenkins & Mitchell, 2009).

One previous study, though, did report an effect of mental state expectedness on RTPJ activation. Saxe and Wexler (2005) manipulated whether mental states were expected or unexpected, relative to the specific background of the protagonist (Saxe & Wexler, 2005). For example, the protagonist of a hypothetical scenario might *want* his or her partner to have an affair – because of an unusual cultural background (e.g., a polyamorous cult) or specific situation (e.g., he or she wanted an excuse to end the relationship). An unusual mental state might then be unexpected in some circumstances but expected in other circumstances. In apparent contrast to the current study, this previous study found that the RTPJ response was enhanced for contextually unexpected beliefs and desires (Saxe & Wexler, 2005). This study suggests that participants constructed a theory (at an explicit or implicit level) about the relationship between the mental state and the protagonist, based on background information provided in the stimulus about the protagonist. Contextually unexpected mental states may require more effortful theory-construction, which may in turn be reflected in more robust recruitment of the RTPJ. By contrast, the stimuli in the current study did not provide any background information about the protagonist; thus, mental states could not be expected or unexpected with respect to a theory about the protagonist. Instead, mental states were merely expected or unexpected with respect to participants' general knowledge about the world. This minimal contrast did not elicit differential recruitment of brain regions for ToM.

The absence of any information at all about the protagonist may also help explain another difference between the current results and previous results: the overall response in all of the ToM brain regions was unusually small for the current stimuli. Whereas in the current experiment, the conditions differed from each other and from rest by approximately 0.1–0.2% of the signal, in previous experiments we have typically found effects of 0.5–0.8% (Saxe & Kanwisher, 2003). We hypothesize that this difference between experiments is a consequence of the stimuli. The current stimuli provided no information about the protagonist's situational context, behavior, or other beliefs and desires – information that could have provoked participants to spontaneously reason about either the origins or the outcomes of the stated belief or desire (Apperly, Riggs, Simpson, Samson, & Chiavarino, 2006; Knobe, 2005; Malle, 1999; Young & Saxe, 2009a): How did the protagonist come to have this mental state? What reasonable behavior follows from this mental state? Instead, the stimuli presented only the target belief or desire – and no other information that would render the belief or desire expected or unexpected, given other mental states or behavior. Future experiments will test the hypothesis that people reason differently (quantitatively or qualitatively) about a protagonist's mental state when rich as opposed to sparse information is provided in the stimulus, concerning the relation between the protagonist and the target mental state.

The current results also inform a recent debate about the role of simulation in thinking about other people's thoughts. That is, to what extent do we understand others' minds by mentally placing ourselves "in their shoes" and simulating their internal experience? On one account, simulation might be easier and therefore more likely when the mind of the target is more similar to one's own mind (Mitchell, Macrae, & Banaji, 2006). For example, it might be easier to simulate the internal experience of a skydiver if you have been skydiving yourself; the earthbound rest of us would have to fall back on drier, cooler "theoretical" inferences about the experience. On an alternative account, simulation might require more effort and therefore more robustly recruit the ToM network when the mind of the target is dissimilar. The current study did not uncover any regions in the ToM network that preferred expected (or similar) to unexpected (or dissimilar) mental states. (The DMPFC did discriminate along this dimension, showing a higher response for unexpected stories, but the higher response for unexpected stories occurred for both mental and physical content.) Thus, the current results do not provide evidence for versions of simulation theory that predict differences for similar versus dissimilar mental states – at least not for the kinds of mental states targeted in the current study, beliefs and desires. Future investigations should focus on how such simulation theories may apply to other mental state content such as stable preferences or attitudes (Jenkins & Mitchell, 2009; Mitchell et al., 2006).

The present findings reveal a role for the RTPJ, LTPJ, PC, and DMPFC in the selective processing of mental states, which cannot

be explained by appeal to high-level attention as manipulated in the current study. These regions were recruited similarly for both unexpected and expected mental states. In addition to supporting a selective role for these regions for mental state reasoning, this pattern suggests that even "irrational" or "unimaginable" mental states count as mental states. Notably, mental states are never "impossible", as physical states may be (Michelon, Snyder, Buckner, McAvoy, & Zacks, 2003; Parris, Kuhn, Mizon, Benattayallah, & Hodgson, 2009). For example, a belief that X can be a false belief (e.g., if X is the proposition: "swimming is a good way for people to grow fins"), but the statement that a person *believes* that X is not necessarily false (i.e. belief attributions are truth-functionally "opaque"). This structural difference between mental and physical representations is worth exploring in future research.

Acknowledgments

The authors gratefully acknowledge Alek Chakroff for his contributions to this project. Scanning was conducted at the Athinoula A. Martinos Imaging Center at McGovern Institute for Brain Research, MIT. This research was supported by the John Merck Scholars program, the David and Lucile Packard Foundation, and the Simons Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuropsychologia.2010.05.012.

References

Apperly, I. A., Riggs, K., Simpson, A., Samson, D., & Chiavarino, C. (2006). Is belief reasoning automatic? *Psychological Science*, *17*, 841–844.

Bledowski, C., Prvulovic, D., Goebel, R., Zanella, F. E., & Linden, D. E. (2004). Attentional systems in target and distractor processing: A combined ERP and fMRI study. *Neuroimage*, *22*(2), 530–540.

Buccino, G., Baumgaertner, A., Colle, L., Buechel, C., Rizzolatti, G., & Binkofski, F. (2007). The neural basis for understanding non-intended actions. *Neuroimage*, *36*(Suppl. 2), T119–127.

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience*, *3*(3), 292–297.

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist*, *13*, 580–593.

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nature Neuroscience*, *3*(3), 277–283.

Ferstl, E. C., Neumann, J., Bogler, C., & von Cramon, D. Y. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, *29*(5), 581–593.

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., et al. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*(2), 109–128.

Friedrich, F. J., Egly, R., Rafal, R. D., & Beck, D. (1998). Spatial attention deficits in humans: A comparison of superior parietal and temporal–parietal junction lesions. *Neuropsychology*, *12*(2), 193–207.

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*(1), 11–21.

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, *19*(11), 1803–1814.

Grezes, J., Frith, C. D., & Passingham, R. E. (2004). Inferring false beliefs from the actions of oneself and others: An fMRI study. *Neuroimage*, *21*(2), 744–750.

Jenkins, A. C., & Mitchell, J. P. (2009). Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex*, *20*(2), 404–410.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, *46*(12), 2949–2957.

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, *9*, 357–359.

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

Michelon, P., Snyder, A. Z., Buckner, R. L., McAvoy, M., & Zacks, J. M. (2003). Neural correlates of incongruous visual information. An event-related fMRI study. *Neuroimage*, *19*(4), 1612–1626.

Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, *18*(2), 262–271.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*(4), 655–663.

Parris, B. A., Kuhn, G., Mizon, G. A., Benattayallah, A., & Hodgson, T. L. (2009). Imaging the impossible: An fMRI study of impossible causal relationships in magic tricks. *Neuroimage*, *45*(3), 1033–1039.

Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: The perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of Cognitive Neuroscience*, *16*(10), 1706–1716.

Perner, J., Aichhorn, M., Kronbichler, M., Wolfgang, S., & Laddurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience*, *1*(3/4), 235–2258.

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*, 59–63.

Posner, M. I., Walker, J. A., Friedrich, F. J., & Rafal, R. D. (1984). Effects of parietal injury on covert orienting of attention. *Journal of Neuroscience*, *4*(7), 1863–1874.

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.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage*, *19*(4), 1835–1842.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, *17*(8), 692–699.

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

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One*, *4*(3), e4869.

Simos, P. G., Basile, L. F., & Papanicolaou, A. C. (1997). Source localization of the N400 response in a sentence-reading paradigm using evoked magnetic fields and magnetic resonance imaging. *Brain Research*, *762*(1–2), 29–39.

Vallar, G., Bottini, G., Rusconi, M. L., & Sterzi, R. (1993). Exploring somatosensory hemineglect by vestibular stimulation. *Brain*, *116*(Pt 1), 71–86.

Vallar, G., & Perani, D. (1986). The anatomy of unilateral neglect after right-hemisphere stroke lesions. A clinical/CT-scan correlation study in man. *Neuropsychologia*, *24*(5), 609–622.

Virtue, S., Parrish, T., & Jung-Beeman, M. (2008). Inferences during story comprehension: Cortical recruitment affected by predictability of events and working memory capacity. *Journal of Cognitive Neuroscience*, *20*(12), 2274–2284.

Vossel, S., Weidner, R., Thiel, C. M., & Fink, G. R. (2009). What is "Odd" in Posner's location-cueing paradigm? Neural responses to unexpected location and feature changes compared. *Journal of Cognitive Neuroscience*, *21*(1), 30–41.

Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporo-parietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 6753–6758.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(20), 8235–8240.

Young, L., Nichols, S., & Saxe, R. (in press). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*.

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *Neuroimage*, *40*(4), 1912–1920.

Young, L., & Saxe, R. (2009a). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*(7), 1396–1405.

Young, L., & Saxe, R. (2009b). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072.