Theory of Mind for you, and for me: behavioral and neural similarities and differences in thinking about beliefs of the self and other.

Hyowon Gweon (hyora@mit.edu), Liane Young (lyoung@mit.edu), Rebecca R. Saxe (saxe@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Abstract

Do we have privileged access to our own mental states, or do we use the same mechanism for thinking about our own minds as we do for thinking about the minds of others? This study featured a task that either induced true and false beliefs in participants or allowed participants to witness another person's true and false beliefs. Later we measured participants' ability to recall their own and others' beliefs, and the recruitment of brain regions for these processes. We found that participants were worse at recalling their own versus others' beliefs, and that brain regions usually associated with ToM tasks were recruited when participants thought about their own beliefs.

Keywords: Theory of Mind; belief attribution; self-reflection; fMRI; RTPJ, LTPJ, DMPFC

Imagine you walk into a coffee shop, order a coffee, and then a minute later pick up someone else's hot chocolate from the counter and start walking out the door. The hot chocolate's rightful owner, Mary, calls out after you, "Why are you taking my hot chocolate?" Presumably you thought you were holding your cup of coffee, and you could generate this explanation, along with an apology, to mollify Mary. But how do you do it? That is, how do figure out what you were thinking, a few moments earlier, when you picked that cup off the counter?

One possibility is that people have direct access to the contents of their own minds, and the reasons for their own actions. Through introspection, people can become directly aware of the beliefs and desires that actually caused their own actions, and retrieve these mental states when explaining or justifying their actions.

An alternative possibility is that people use a 'Theory of Mind' to infer the beliefs and desires that most likely caused their own actions. Imagine the scenario were reversed: you've just ordered a hot chocolate, and Mary, who ordered a coffee, picks up your cup of hot chocolate and starts to walk off. In this situation, most adults can infer Mary's false belief; this inference allows people to recognize Mary's mistake, and not blame her for hot chocolate theft. Young children, by contrast, see only that Mary is taking their hot chocolate, and say that Mary must be a mean person (Fincham & Jaspers, 1979).

Do people reason about their own past beliefs by direct introspection, or by applying a Theory of Mind, relying on the same mechanism that supports reasoning about the minds of others? These alternative hypotheses can be tested behaviorally and neurally. Behaviorally, if people use direct introspection to recall their beliefs, we might expect that reasoning about one's own beliefs would be more accurate than reasoning about others' beliefs. By contrast, if people have to infer their own past beliefs, using the same Theory of Mind, then they might make the same mistakes, whether reasoning about their own or others' beliefs (Saxe, 2005). Indeed, people might be even worse at reasoning about their own beliefs than about others' beliefs. People usually act on their own beliefs without representing them *qua* beliefs (Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000). That is, at the moment of taking the hot chocolate (which you believe is your coffee) you are unlikely to explicitly attribute to yourself a belief, i.e. "I believe this is my coffee". Belief attributions to the self occur only rarely, when the beliefs one acts on turn out to be false, or the actions have negative consequences. Thus, there may be an asymmetry between ToM for ourselves and for others: we often need to explain others' actions using ToM, but not as frequently to explain our own.

Developmental evidence favors the second alternative: children learn to reason about their own past false beliefs at the same time that they learn to reason about others' current false beliefs (Atance & O'Neill, 2004; see Wellman, Cross, & Watson (2001) for a review). In these experiments, children see a crayon box (and form the belief that the box contains crayons), but the box is shown to contain candles. In the third person version, children are asked: "when another child comes into the room, and first sees the box, what will she think is inside?" Five year olds understand false beliefs, and say "crayons"; three year olds don't, and say "candles". In the first person version, before the candles are revealed, the children are induced to act on their false belief (i.e., to get a piece of paper to draw on with the crayons). After seeing the candles, children are then asked: "why did you get the piece of paper?" Five-year-olds say, "because I thought there were crayons in the box". Three-year-olds, however, do not appeal to their own prior beliefs but refer to irrelevant facts that occurred after the action (e.g., there were candles in the box) or confabulate other reasons (e.g., the paper was the floor).

We can also test whether belief attribution to self relies on Theory of Mind by identifying which brain regions are recruited when people recall their recent beliefs in order to explain their own actions. Many neuroimaging studies have investigated the brain regions that people use when thinking about someone else's false beliefs (Saxe & Kanwisher, 2003; Perner, Aichorn, Kronblicher, Staffen, & Ladurner, 2006; Gallagher et al., 2000). Remarkably, these neuroimaging studies have converged on a distinct network of regions including the right and left temporo-parietal junction (TPJ), the precuneus (PC), and regions in the medial frontal cortex (MPFC). To our knowledge, however,

no fMRI study has directly compared reasoning about one's own beliefs to reasoning about another person's beliefs. Although some neuroimaging studies have compared thinking about the self to thinking about others, these studies asked participants to reflect on stable personality traits (Jenkins, Macrae, & Mitchell, 2008) or current affective states (Ochsner et al., 2004), or to read stories that require ascriptions of beliefs to themselves in hypothetical situations (Vogeley et al., 2001). A straightforward comparison between ToM for the self and others should require participants to act on beliefs, or watch others act on the same true and false beliefs, based on the same evidence, and then to reason about those beliefs in matched circumstances.

Thus, the current study addressed the following questions. First, do people have privileged access to their own past beliefs, such that, behaviorally, they are more accurate in recalling their own beliefs versus others' beliefs? Second, is the neural mechanism that has been shown to support ToM for others also recruited for tasks that involve thinking about one's own thoughts? We devised a task that naturally induced true and false beliefs in the participants. Participants were then shown whether they were right or wrong, and finally instructed explicitly to think back to their prior true and false beliefs. A different set of subjects participated in an analogous task, using the same stimuli and instructions but targeted another person's beliefs; participants watched another person act on true and false beliefs, and then later thought back to that person's beliefs.

Experiment

To assess the behavioral and neural differences in how people think about their own versus others' beliefs, we designed a task that leads participants to either: (a) generate a false or true belief about images or (b) encounter another person's false or true belief about the same images. Then, 40 – 50 minutes later, we asked participants to judge: (a) whether they were right / wrong about the images or (b) whether the other person was right / wrong about the images. We measured participants' recall accuracy for their own and others' past beliefs. In addition, participants completed these tasks inside an fMRI scanner so that we could also measure neural activity while people thought about their own or others' beliefs.

Methods

Participants Twenty-four healthy adults (18 – 25 years, 8 males) participated in the experiment. Twelve participated in the "Self" version, and the other twelve participated in the "Other" version of the experiment. All participants were native speakers of English, right-handed, and had normal or corrected-to-normal vision.

Stimuli Forty-eight hand-drawn color drawings were used. Thirty-six of these pictures were presented both as a whole picture (Whole Picture) and partially occluded to reveal only a small part of the picture (Part Picture). The Part

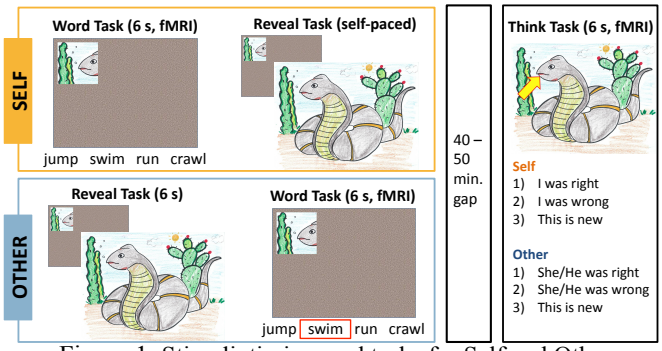

Figure 1. Stimuli, timing, and tasks for Self and Other versions of the experiment. The Part Picture shown here deliberately misleads people to form a false belief.

Pictures of some drawings were deliberately designed such that the participant would be misled about the object's identity (e.g., the visible part in the Part picture looks like a fish, but it is actually a snake in the Whole picture, see Figure 1). Some Part Pictures provided an accurate representation of the object in the picture such that the participant would form a true belief, while others provided insufficient information about the object. The remaining 12 drawings were presented only as Whole Pictures, in the last part of the experiment to serve as control "new" drawings. All stimuli were presented in Matlab (R2010a) using Psychtoolbox 3 (http://psychtoolbox.org).

Procedure – Self The experiment consisted of three different tasks: the Word Task, the Reveal Task, and the Think Task. In the Self version of the experiment, participants completed the Word Task first, followed by the Reveal Task and the Think Task. In the Word Task, participants were instructed to look at the Part picture with four words presented at the bottom of the screen, and to choose the word they thought was most closely associated with the hidden picture. Of the four words, one was always the "correct" answer, which was associated with the fully revealed picture (i.e., Whole Picture). For the pictures that were deliberately misleading, one of the word choices was a "lure" word, which was associated with the false belief the participants would generate if they were misled by the Part picture. Other words were fillers that were not associated with either picture version (i.e., Part or Whole). Participants were instructed to select a word if they could not figure out the content of the picture. The Word Task was divided into two runs (18 trials in each run, 36 trials total). Each trial was 6 seconds long, with 10 seconds fixation.

In the Reveal Task, participants saw the 36 Part pictures, and were instructed to press a button to reveal the Whole picture. Therefore, after each button press, they could see what each drawing really depicted. This Reveal Task was self-paced (no fMRI data were collected during this task.)

Then, participants completed tasks for a different study for 40 – 50 minutes before the final task. One of the tasks was a functional localizer designed to identify the ToM network in each individual's brain. People read stories that required inferences about a character's beliefs with stories

that required inferences about a physical representation (e.g., an outdated map or a photograph). Details of this localizer task can be found on the SaxeLab website (http://saxelab.mit.edu/superloc.php).

In the final Think Task, participants saw 12 new images in addition to all 36 images they had seen in the previous two tasks (e.g., Word, Reveal). In each image, an arrow pointed to the main object in the image. The participants were instructed to think back to what they thought about the object during the Word task, and to choose one of the following response options: (1) I was *right* (about the identity of the object in the picture), (2) I was *wrong*, and (3) This is *new*. The Think task was also divided into two runs, with 24 trials in each run. Each trial was 6 seconds long, during which the picture remained on the screen, followed by 10 seconds fixation. Participants could respond as long as the picture remained on the screen.

Procedure – Other In the Other version of the experiment, the ordering of the tasks reflected a fundamental difference between thinking about one's own versus others' beliefs: sometimes we already know the true state of the world when we observe others' actions. Therefore, participants in the Other version first completed the Reveal task. Each Part picture was presented for 3 seconds, and then the Whole picture was revealed. Participants were instructed to press a button when the picture changed from Part to Whole[1].

Then participants completed the Word task with different instructions. Participants were told that a second participant (who had not yet seen the Whole pictures) would perform the Word task and choose one of the four words that he or she thinks is the most closely associated with the hidden Whole picture. Participants were told that this person's response would be projected to the participant's screen (e.g., as a pink square around the chosen word). The participant's task was to press the same button that the other person had pressed to ensure that participants encoded the other person's response. In fact, there was no 'second participant'; the other responses were generated by a computer. The 'second participant' chose the correct word on 12 trials (with informative Part pictures), the lure word on 12 trials (with misleading Part pictures), and one of the other words on 12 trials (with uninformative Part pictures). The picture remained on the screen for 6 seconds, and the pink square (representing the second participant's choice) came up 3 seconds after the onset of each picture.

Finally, after 40 – 50 minutes, participants completed the Think task, again with different instructions. They were told to think back to the second participant's belief about the drawing, and to choose one of the following response options: (1) She/He was right (about the identity of the object in the picture), (2) She/He was wrong, and (3) This is new. The timing and the number of the trials were the same as the Self version of the task.

Behavioral data analysis Participants' button responses and RT during the Word Task and the Think Task were collected and analyzed to determine the judgment accuracy and speed in the Think Task. In the Self version, participants constructed their own beliefs about the picture during the Word Task. Therefore, judgment accuracy of participants in the Self version during the Think Task was determined relative to each participants' own word choices during the Word task. For example, if the participant chose the word 'swim' in the Word Task (a lure answer for this misleading drawing; see Figure 1) but chose "I was right" during the Think Task, this judgment was considered inaccurate, as the participant had previously formed a *false* belief about the picture. In the Other version, participants always saw the other person making an incorrect choice when the drawings were deliberately misleading or ambiguous, and always a correct choice if the Part Picture provided enough information about the drawing; therefore, participants' accuracy during the Think Task was based on these pre-determined word choices. Behavioral data for one of the participants was lost due to experimental error, and therefore excluded from further analysis.

fMRI data collection and analysis Participants were scanned on a 3T Siemens scanner at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted structural images were collected in 256 saggital slices (TR = 2.53s, TE = 3.39ms, flip angle = 9.0°) with 1.0 mm isotropic voxels. Functional data were acquired in 3.1 x 3.1 x 4 mm voxels in 64 interleaved near axial slices covering the whole brain, using standard echoplanar imaging procedures (TR = 2 s, TE = 30 ms, flip angle = 90°). These sequences used prospective acquisition correction (PACE), which adjusts the slice acquisitions during scanning to correct for head movement up to 8 degrees and 20 mm.

fMRI data were analyzed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and custom software written in Matlab. Each participant's data were off-line motion corrected and then normalized onto a common brain space (Montreal Neurological Institute (MNI) template). Data were then smoothed using a Gaussian filter (full width half maximum = 5mm). All functional images that exceed a scan-to-scan motion threshold of 1.5mm and Z-score of 3 in global intensity were regressed out using the Artifact Detection Tool (ART). The mean number of images excluded for each participant was 40.2 (SD = 51.2, 4.2% of all images) for the Self group, and 25.5 (SD = 30.9, 2.6%) for the Other group ($p = ns$). The experiment was modeled using a boxcar regressor. An event was defined as presentation of an image that participants responded with "Right", "Wrong", or "New". Data were high- pass filtered during analysis (cutoff 128 seconds).

[1] Note that the Reveal Task for participants in the Self condition was self-paced. To address the potential concern about participants in the Self condition having less (or more) exposure to the whole picture during the Reveal Task, we recruited a separate group of participants just for the behavioral part of the Self version. The behavioral results mirrored the pattern found in the Self group reported here.

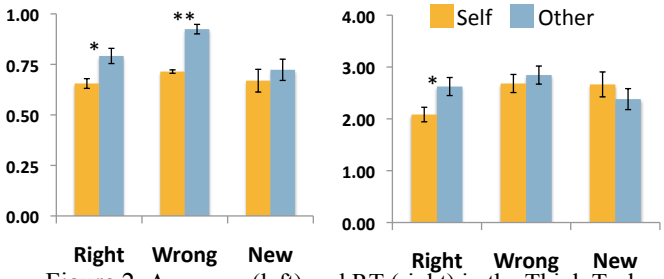

Figure 2. Accuracy (left) and RT (right) in the Think Task. (*: $p < 0.05$, **: $p < 0.001$)

Both individual ROI (Region of Interest) and whole-brain analyses were conducted, separately for participants in the Self (N=12) and Other (N=12) versions. In the whole-brain analyses, the false-positive rate was controlled at $\alpha < 0.05$ (corrected) by performing Monte Carlo permutation tests using the SnPM toolbox for SPM5 (http://www.sph.umich.edu/ni-stat/SnPM/) to empirically determine the voxel-wise t and cluster size (k, contiguous voxels) thresholds. Three functional ROIs, the TPJ bilaterally and DMPFC, were defined for each participant individually from the Belief versus Photo contrast of the localizer task. The RTPJ was defined in all 24 participants, LTPJ and DMPFC in 22 participants. ROIs were defined as contiguous voxels active at a threshold of $p < 0.001$, uncorrected, $k > 10$. For each ROI, we report the average percent signal change (PSC) of the raw BOLD signal in each condition[2]. For the purposes of statistical analyses, we averaged PSC across the time points during which the pictures were presented (4 – 10 seconds after the image onset, to account for hemodynamic lag) to obtain a single PSC value for each region in each participant.

Results

Behavioral Results

Preliminary analysis of the Word Task responses confirmed that the drawings successfully induced false and true beliefs in the Self participants: participants chose the correct and incorrect word choices in 48.5% and 51.5% of the 36 trials, respectively.

Our main goal was to see whether people are more accurate, less accurate, or no different, in recalling their own previous beliefs (e.g., true or false beliefs) as compared to other people's beliefs. We found that the average judgment accuracy during the Think Task was lower for Self than Other. When people reported prior true beliefs (e.g., "I was right" or "She/He was right"), participants were less accurate when recalling their own (66%) versus another's belief (79%, $z = 2.35$, $p < .05$, Mann-Whitney test, see Figure 2). They were also less accurate in reporting their own prior false beliefs (71%)(i.e., "I was wrong") than others' false beliefs (i.e., "She/He was wrong"; 92%, $z = 3.77$, $p < .001$). However, there was no difference in accuracy when people judged a picture as new (67% (Self) vs. 72% (Other), $z = 0.7$, $p = ns$). Overall RT showed no difference between Self and Other groups (2.48 (Self) vs. 2.62 (Others), $t = 0.59$, $p = ns$), but people in the Self group were faster to judge that they were "Right" than people in the Other group (2.08 (Self) vs. 2.62 (Other), $t = 2.36$, $p < 0.05$).

These results suggest that people are not in fact better at recalling their own beliefs. On the contrary, they were worse at recalling their own versus others' beliefs. Importantly, this difference was not due to participants in the Self group consciously or unconsciously "lying" to inflate their accuracy: participants in the Self group were no more likely to inaccurately report "I was right" when they actually gave an incorrect answer in the Word Task (31.7% of "I was right" responses), than to inaccurately report "I was wrong" when they actually chose the correct answer in the Word Task (27.9% of "I was wrong" responses, $t(10) = 1.41$, $p = 0.19$). Instead, people seem to be genuinely worse at accurately recalling the beliefs upon which they acted. Participants in the Other group also did not differ in their tendency to respond that the other person was right when they in fact were wrong, and to report that the other was wrong when they in fact were right (10.1% vs. 6.9%, $t(12) = 1.1$, $p = ns$).

fMRI Results

We asked whether regions in the ToM network, which show robust and selective activation when people think about other people's thoughts and beliefs, are also recruited when people think about their own past thoughts and beliefs. We predicted that participants in both the Self and Other groups would show heightened response in these areas when they indicated "I (She/He) was right" (true belief) or "I (She/He) was wrong" (false belief), than when they indicated that "This picture is new".

The whole brain analysis for the Right & Wrong vs. New contrast confirmed that this was indeed the case in the Self group: we found bilateral TPJ and MPFC activation (see Figure 3). By contrast, we found a very different pattern in the Other group: left inferior frontal gyrus (IFG) / rostrolateral prefrontal cortex (RLPFC), middle frontal gyrus (MFG) bilaterally, left inferior parietal lobule (IPL), superior and medial frontal gyrus, which are brain regions commonly associated with non-spatial working memory tasks (D'Esposito, Postle, & Rypma, 2000) or higher-order mental operations such as relational reasoning (Christoff, Ream, Geddes, & Gabrieli, 2003). To take a more detailed look at the response profiles of these regions, we identified bilateral TPJ and MPFC in individual participants from a functional localizer scan (see Methods for details). The average PSC values for each trial type (sorted by response, "I(He/She) was right", "I(He/She) was wrong", "This is

[2] PSC was calculated by first extracting the average BOLD magnitude of the ROI in each condition for each time point after the onset of the stimulus, then subtracting the baseline (average BOLD magnitude of the ROI during fixation) from these values, and divided this with the baseline BOLD (PSC(condition,time) = 100* (Resp(condition,time) – baseline) / baseline). The result is a timecourse showing the percent signal change relative to baseline for each condition at each time point.

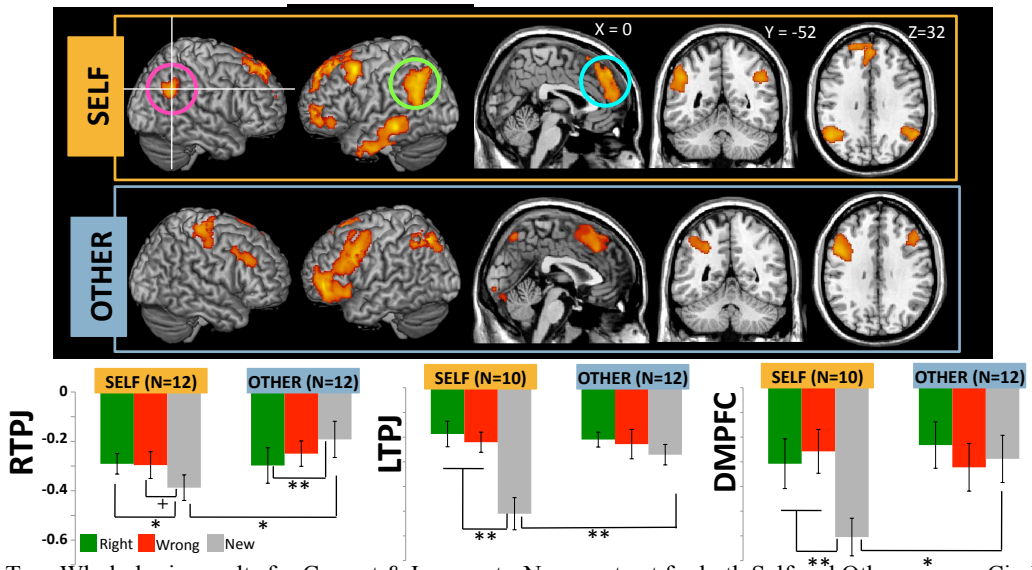

Figure 3. Top: Whole-brain results for Correct & Incorrect - New contrast for both Self and Other groups. Circles indicate the brain regions that were identified in individual participants: RTPJ (pink), LTPJ (light green), and DMPFC (cyan). Bottom: Average PSC values for each of the ROIs during Right, Wrong, and New responses. (+: $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$)

new") from these ROIs were entered into a repeated measures ANOVA with Group (Self, Other) as a between-subjects factor and Response (Right, Wrong, New) as a within-subjects factor (see Figure 3).

In the RTPJ, we found a significant interaction between Group and Response ($F(2,44) = 6.81$, $p < .005$, partial $\eta^2 = .24$): planned comparisons revealed that there was a significant difference between Right and New ($t(11) = 2.72$, $p < .05$) and a marginally significant difference between Wrong and New in the Self group ($t(11) = 2.15$, $p = .055$). That is, the activity in the RTPJ was higher when people reported that they were Right or Wrong than when they judged pictures as New. However, there was a reverse trend in the Other group: activity was higher for New than Right responses ($t(11) = -3.36$, $p < .01$). All other differences within the Other group were not significant.

A similar trend was found in the other two ROIs. In the LTPJ and DMPFC, we found a main effect of Response (LTPJ: ($F(2,40) = 17.72$, $p < .001$, partial $\eta^2 = .47$), DMPFC: ($F(2,40) = 12.37$, $p < .001$, partial $\eta^2 = .38$)) and an interaction between Group and Response (LTPJ: ($F(2,40) = 8.79$, $p = .001$, partial $\eta^2 = .31$) DMPFC: ($F(2,40) = 12.19$, $p < .001$, partial $\eta^2 = .38$). Again, these were driven by the difference in Right vs. New (LTPJ: $t(9) = 4.50$, $p < .001$, DMPFC: $t(9) = 3.98$, $p < .005$) and Wrong vs. New (LTPJ: $t(9) = 5.12$, $p = .001$, DMPFC: $t(9) = 8.76$, $p < .001$) in the Self group. Both LTPJ and DMPFC showed higher activity when people said they were Right or Wrong than when they said the picture was New. In the Other group, we found no difference between the three responses.

Finally, we compared activity during the New responses between the Self and Other groups. Results showed that the activity during the New response was significantly higher in the Other group than in the Self group, in all three ROIs (RTPJ: $t(22) = -2.31$, $p < 0.05$, LTPJ: $t(20) = -3.20$, $p = .005$ DMPFC: $t(20) = -2.57$, $p < 0.05$).

Overall, these results suggest that regions in the ToM network – bilateral TPJ and DMPFC – are recruited when people think about their own prior beliefs.

Discussion

The current study allowed us to directly compare the cognitive and neural aspects of ToM for ourselves and ToM for others. We experimentally induced the "Self" participants to act on true and false beliefs and then later asked them to recall those beliefs. The "Other" participants saw another person acting based on his or her true or false beliefs, and then recalled that person's beliefs. We compared the behavioral performance and neural activity between the Self and Other participant groups.

First, we found that people are worse at remembering their own past beliefs (whether they were true or false) than remembering another person's past beliefs, contrary to the hypothesis that people have privileged access to their own (past) mental states. Second, when people reflected upon their past beliefs, compared to when they simply judged whether they had seen the picture, we observed enhanced activity in key regions for ToM, the RTPJ, LTPJ, and DMPFC. These results suggest that when people think back to their own (recent) beliefs as explanations for their own actions, the same Theory of Mind mechanisms are recruited as when people explain and predict others' actions.

Does this finding suggest that people do not have *any* privileged access to the contents of their own minds? The strongest version of this hypothesis predicts that people must always infer their own thoughts by observing their own actions (Bem, 1972): when sitting quietly in a room with someone else, people would know as little about their own thoughts as about the other person's! We do not endorse this strong view. On the contrary, we suggest that people use different mechanisms for experiencing their own current perceptual and epistemic states, versus inferring and attributing others' current, and anyone's past, mental states.

As a consequence, there is an asymmetry in when people think about their own beliefs versus others' beliefs: ToM is frequently used to understand other people's past, current and future actions, and also (but relatively rarely) used to explain one's own past actions.

Our data also provide evidence against the claim that brain regions for ToM are recruited for resolving conflicts between false representations and reality (Sommer et al., 2007) or for low-level attentional processes invoked by false belief reasoning (Mitchell, 2008). True and false belief responses elicited equally high activity in the RTPJ, LTPJ, and DMPFC. Typically, when people act on true beliefs, they can explain their behavior based on reality alone; however, the current task explicitly required participants to think about their true and false beliefs alike. These results suggest that ToM brain regions are recruited for thinking about true and false beliefs – one's own and other people's.

One unexpected result was the lack of a neural difference between Right/Wrong versus New responses in the Other group: instead, the neural activity during New responses was just as high as during the other two responses (Right/Wrong). One possible account is that participants in the Other group engaged in ToM for all conditions, including when they were reporting that a New picture hadn't been seen by the other person. Consistent with this account, we found a higher response in ToM brain areas for people who responded that another person had not seen a picture before, compared to people who responded that they themselves had not seen a picture before. To recognize something as new or familiar, we simply need to introspect on our current experience. However, to report the current feeling of familiarity in another person, we may need to think about that person's previous experience or belief. If this were indeed the case, the fact that the participants did not simply use their own experience to decide whether the picture is new (since pictures that were new to the other person were also new to the participants themselves) raise an interesting question about the spontaneous and automatic engagement of ToM in social, interpersonal contexts, versus the conservative use of ToM for the self.

To our knowledge, the current study represents the first attempt to directly compare belief attribution to the self versus other. The results suggest important asymmetries in how and when we think about our own beliefs, resulting in lower accuracy for retrieving and representing one's versus others' own beliefs. The neural results suggest that when prompted to think about our own beliefs, we rely on the same neural network for ToM as we do for representing the beliefs of others.

Acknowledgments

Thanks to Jacquie Pigeon and Hannah Pelton for help with data collection. This research was funded by a John Merck Scholars Grant.

References

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed)., Advances in Experimental Social Psychology, Vol. 6, 1 – 62. New York, NY: Academic Press.

Christoff, K., Ream, J., Geddes, L., & Gabrieli, J. (2003). Evaluating self-generated information: anterior prefrontal contributions to human cognition. *Behavioral Neuroscience, 117*(6), 1161-1167.

D'Esposito, M., Postle, B., & Rypma, B. (2000). Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. *Experimental Brain Research, 133*(1), 3-11.

Fincham, F. D., & Jaspers, J. (1979). Attribution of responsibility to the self and other in children and adults. *Journal of Personality and Social Psychology, 37*(9), 1589-1602.

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.

Jenkins, A., Macrae, C., & Mitchell, J. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences, 105*(11), 4507.

Malle, B.F., Knobe, J., O'Laughlin, M.J., Pearce, G.E., & Nelson, S.E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions. *Journal of Personality and Social Psychology, 79*(3), 309-326.

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

Ochsner, K., Knierim, K., Ludlow, D., Hanelin, J., Ramachandran, T., Glover, G., et al. (2004). Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience, 16*(10), 1746-1772.

Perner, J., Aichorn, M., Kronblicher, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: the roles of right and left temporo-parietal junction. *Social Neuroscience, 1*(3-4), 245-258.

Saxe, R. (2005). Against simulation: the argument from error. *Trends in Cognitive Sciences, 9*(4), 174-179.

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.

Sommer, M., Dohnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. *Neuroimage, 35*(3), 1378-1384.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., et al. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage, 14*(1 Pt 1), 170-181.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development, 72*(3), 655-684.