Theory of Mind brain regions are sensitive to the content, not the structural complexity, of belief attributions

Jorie Koster-Hale (jorie@mit.edu) and **Rebecca R. Saxe (saxe@mit.edu)**
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

A distinct group of brain regions, the 'Theory of Mind (ToM) network', is implicated in representing other people's mental states, yet we currently know little about which aspects of mental state attribution are represented or processed in these regions. Using fMRI, we investigated whether ToM regions, compared to language-processing regions, are sensitive to two dimensions along which mental state attributions vary: (1) structural complexity and (2) social content of the attributed thought. In short vignettes describing a character's belief, the belief structure was either *first-order* or *higher-order*, and the content was *mundane* or *socially-relevant*. All ToM regions showed sensitivity to distinctions in content; no ToM region showed sensitivity to structural manipulation. By contrast, language regions were sensitive to both manipulations. We conclude that while increased structural complexity of belief attributions modulates language processing, this type of complexity is not part of the representational space of the ToM-network.

Keywords: Theory of Mind; False Belief, Language, fMRI

Introduction

Mental state attribution exists in a very rich conceptual space – without much effort, we can ascribe a variety of mental states to other people, and make quick and subtle judgments about them. Moreover, we can easily characterize a mental state along a number of dimensions, such as who holds it, what kind of mental state it is (e.g. a belief, desire, or doubt), what the belief is about, how reasonable we find it, whether the content is relevant to our own lives, and how probable it is that it will be believed next week.

Yet despite the range and flexibility of these inferences, mental state attribution gives rise to a surprisingly uniform neural response. A specific set of regions, often called the Theory of Mind network, consisting canonically of the bilateral temporo-parietal junction (TPJ), right superior temporal sulcus (rSTS), medial precunius (PC), and medial prefrontal cortex (MPFC), shows robust and systematic response to a variety of stimuli that invoke a mental state attribution, including stories and cartoons (Fletcher et al. 1995; Goel et al. 1995; Gallagher et al. 2000, 2002; Mitchell et al. 2002; Saxe and Kanwisher 2003; Perner, Aichhorn, Kronbichler, Wolfgang, & Laddurner, 2006; Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007; Van Overwalle 2009, Walter et al 2010).

This combination of cognitive flexibility coupled with a robust and seemingly invariant neural response provides chance to examine the mapping between the neural response and final cognitive product: though we currently know very little about which aspects of mental state attribution are represented or processed in theory of mind regions, or what that representation looks like, we have the means to manipulate the cognitive representation at a fairly high level, and a precise place to look for changes in the neural representation.

Thus, to begin answering these questions, we investigated the extent to which brain regions involved in theory of mind processing show sensitivity to features that vary within the space of mental state attribution. We asked whether ToM regions are sensitive to two broad dimensions along which mental state attributions can vary: (1) the structural (or syntactic) complexity and (2) the content of the attributed belief. We manipulated structural complexity by manipulating the first versus higher-order status of the belief – a manipulation that has often been employed to increase the difficulty of ToM tasks. We manipulated the content of the belief by varying its the social relevance.

As well as varying features within the space of belief attribution, these manipulations vary along linguistic dimensions – saliency and syntactic complexity. Thus, to serve as a comparison, we asked whether high-level (sentence-level) language processing regions show sensitivity to these manipulations, and if so, whether the response profile in the ToM regions differed from the response profile in the language regions.

We tested each of these possibilities in two steps. First, we used a functional localizer to identify language processing and Theory of Mind regions within the same set of individuals (Experiment 1). Second, we examined the effect that our manipulations of the structure and content of belief attributions had on the brain regions implicated in language and ToM (Experiment 2).

Methods

Participants:

Twenty naïve right-handed adults (aged 21-44, mean 27; 15 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. All 20 participants did both Experiment 1 and Experiment 2 in a single scan session. Two participants' data were excluded due to excessive movement.

Stimuli and Design:

Experiment 1: ToM and Language Localizer

The stimuli consisted of 24 short stories and 12 lists of non-words. Twelve of the stories were described a situation in which someone held a *false belief*, e.g.:

After going to the gym, Kevin returned to his new apartment, which he had just recently moved into. He got upstairs and threw off his sweaty clothes, ready for a hot, steaming shower. Regrettably,

Kevin's roommate had thrown out an important note from the plumber, so Kevin didn't know that the pipes had broken and were currently full of cold pond water.

The remaining 12 stories were *false photograph* stories, describing a situation in which there was a false physical representation of the world, such as an out-of-date photograph or advertisement, e.g.:

Last week, many fliers and signs were posted, advertising the open house of an apartment building that had just been built downtown. The ads had pictures of the granite counters, the balconies, the huge swimming pool, the art gallery, and the gym. Regrettably, before the open house, the building caught fire, and today's paper reported that most of the building was destroyed.

Both of these kinds of story require the reader to deal with incorrect or outdated representations of the world, and so are similar in their meta-representational and logical complexity; however, they differ crucially in whether the reader is building a representation of someone else's mental state, and thus comparing them serves to localize those regions recruited particularly for processing mental states. See Saxe and Kanwisher (2003) and Dodell-Feder et al (2010) for further discussion.

To control for low-level linguistic properties and possible processing confounds, the conditions were additionally matched for number of words, number of syllables per word, Flesch reading ease, number of noun-phrases per sentence, lexical frequency, log-transformed lexical frequency, number of negations, and general syntactic form (e.g. number of relative clauses), all $p >> 0.1$

From each matched pair of stories, a *word-list* was created, consisting of a random subset of the unique words from each story. A matched *non-word* list was created by selecting legal bigram combinations that were matched to each word on length, number of syllables, and bigram frequency.

Processing pronounceable non-words engages many of the low-level processes required for (visual) language processing, such as visual processing, phonological recognition and composition, and working memory, without recruiting higher-level processes, such as lexical access, word and sentence level composition, syntactic structure building, or semantic computation. Processing sentences, on the other hand, engages both low-level visual and phonological processing and also higher-level linguistic processes. Thus, comparing sentences to lists of pronounceable non-words serves to localize those regions specifically recruited for language processing on the word and sentence level (Fedorenko et al 2010, Cutting et al. 2006; Friederici et al. 2000; Hagoort et al. 1999; Heim et al. 2005; Humphries et al. 2006, 2007; Indefrey et al. 2001; Mazoyer et al. 1993; Petersen et al. 1990; Vandenberghe et al. 2002).

All stimuli were presented one word at a time (screen center) for 350 ms each, following the procedure of Fedorenko et al (2010). At the end of each story/list of non-words, a probe word was presented for a 2s answer period. Participants were asked whether the probe word appeared in the preceding story/list (a match-to-sample task): 50% of the probes were matches and 50% were novel, drawn from an unseen stimulus in the same condition. Participants were also told to read the stories for content, and asked to think about and visualize the scene. Trials were separated by 12-18 seconds of fixation. The text of each story was presented in a white 40-point font on a black background, using Matlab 7.10 running on an Apple MacBook Pro, and the order of conditions was counterbalanced across runs and participants.

Experiment 2: Structure and Content

The stimuli consisted of 40 short vignettes, which introduced two people and a context, and then described a belief that one of the characters held. The content of the belief itself was either *mundane* (thoughts about e.g. housework, haircuts, paint colors) or *socially relevant* (e.g. scandal, drugs, sexual relations). The form the belief description was either *first-order* ('John thinks that …') or *higher-order* ('John thinks that Mary suspects that he knows that…'). Each vignette appeared in all four conditions, counter-balanced across participants; each participant saw one of the four versions of each vignette for a total of 40 stories, e.g.:

*Jessica was just hired as the new program director at a local non-profit that works on [*1st: *raising money and recruiting volunteers for] special needs education.*

*When, Steve, the on-site manager, met her, he was very impressed, mostly because he thinks that [*higher: *Jessica suspects that he believes that] …*

*[*mundane: *as a trained negotiator and long-time networker, Jessica will be successful at bringing in new grant money.]*

*[*social: *as a very attractive and large-chested woman, Jessica will be successful at bringing in new grant money.]*

Each condition was matched for word count, number of syllables per word, Flesch reading ease score, number of noun-phrases), lexical frequency, log-transformed lexical frequency, and general syntactic form (e.g. number of relative clauses), such that there was no significant difference between *mundane* and *socially-relevant* stories or between *first-order* and *higher-order* ones (all $p >> 0.1$).

Stories were presented in a pseudo-randomized order, with the order of conditions counterbalanced across runs and participants. Full stories were presented all at once for 20 seconds, followed by 12 s of fixation on a black screen. 10 stories were presented during each of four runs for a total run time of 22 min and 56 seconds. Each story was presented in a white 40-point font on a black background, using Matlab 7.10 running on an Apple MacBook Pro. Participants were asked to press a button when they were done reading.

Participants were informed that after the scan there would be a memory task; they were told to not try to memorize the details of the stories, but to read the stories as fully and deeply as possible, as if they were reading a novel.

After scanning, participants were presented with a self-paced memory task in which they saw the same 40 stories that they saw in the scanner, presented in a pseudo-randomized order. Half of the stories (distributed evenly across conditions) were slightly modified, with a change to e.g. one of the main character descriptions, the location of the story, or the belief content. Participants were asked to determine whether this version of the story was the same version that they saw in the scanner.

fMRI Data Acquisition

fMRI data were collected in a 3T Siemens scanner at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, using a 12-channel head

coil. Using standard echoplanar imaging procedures, we acquired blood oxygen level dependent (BOLD) data in 30 near axial slices, using 3 x 3 x 4 mm voxels (TR = 2 s, TE = 30, flip angle = 90◦). To allow for steady state magnetization, the first four seconds of each run were excluded.

Data processing and analysis was performed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. The data were realigned, normalized onto a common brain space (Montreal Neurological Institute (MNI) template), spatially smoothed using a Gaussian filter (5 mm kernel) and subjected to a high-pass filter (128 Hz).

fMRI Analysis:

Both experiments were modeled using a boxcar regressor, convolved with a standard hemodynamic response function (HRF). The general linear model was used to analyze the BOLD data from each subject, as a function of condition. The model included nuisance covariates for run effects, global mean signal, and an intercept term.

Exp. 1: ToM and Language Localizer

A second-level random effects analysis was performed on the contrast images generated for each individual to identify brain regions showing reliable differences between belief and photo stories ('Theory of Mind' regions) and between photo stories and nonwords ('Language' regions; thresholded at $p<0.001$, uncorrected, $k>10$). Based on the results of the whole-brain analysis, functional regions of interest (ROIs) were defined for each individual, as a set of at least 10 contiguous voxels that showed a significant difference between conditions (thresholded at $p<0.001$, uncorrected).

To measure the response of these ROIs to the localizer stimuli without the bias of non-independent data, we used a cross-validation technique. Individual subject ROIs were defined using two runs of data, and the response was extracted from the excluded, independent run. This process was iterated over all three runs, allowing us to calculate, in each of the individual regions of interest (ROIs) defined using the localizer, the average percent signal change (PSC) relative to baseline for each time point in each condition, averaging across all voxels in the ROI and across all blocks in the condition, where PSC(t) = 100 × (average BOLD magnitude for condition (t) – average BOLD magnitude for fixation) / average BOLD magnitude for fixation. We averaged the PSC across the entire presentation – offset 6s from presentation time to account for hemodynamic lag – to get a single PSC for each condition, in each ROI, in each participant (Poldrack, 2006). These values were then averaged across subjects to get a PSC value for each condition for each ROI.

Exp. 2: Structure and Content

In each of the individual regions of interest (ROIs) defined using the localizer, we calculated the average percent signal change (PSC) as in Experiment 1, to get a PSC value for each condition for each ROI.

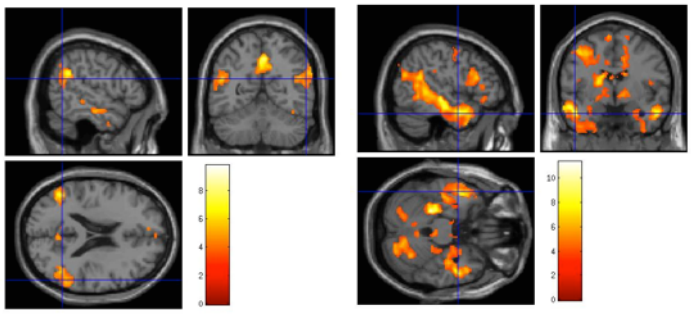


Figure 1a,b: Functional localizer results – ToM (left) and Language (right). Brain regions in which the bold signal was higher for stories about mental representations compared to stories about physical representations; and for physical representation compared to non-words (N = 18, random effects analysis, p < 0.001, uncorrected.)

Results and Discussion

Experiment 1: ToM and Language Localizer

Theory of Mind regions:

A whole brain random effects analysis revealed five main regions that showed greater activation for *false belief* stories compared to *false photograph* stories (uncorrected, $p < 0.001$, $k > 10$): right and left temporo-parietal junction, right superior temporal sulcus, medial precuneus, and dorsal medial prefrontal cortex. Identifying the set of brain regions that are considered a core part of the Theory of Mind network, these results replicate a number of studies using a similar functional localizer task (e.g. Saxe & Kanwisher, 2003). These ROIs were then identified in each individual, using the same threshold: RTPJ (identified in 18/18 individuals), LTPJ (15/18), RSTS (14/18), PC (15/18), and DMPFC (12/18), (Figure 1a).

The results from the cross-validation were analyzed using pair-wise comparisons of the response to *false belief* stories, *false photograph* stories, and *non-word* lists. Paired-sample t-tests revealed that all of the individually localized ToM regions show a significant difference between *false belief* and *false photograph* (all $p<.05$), and between *false belief and non-words* (all $p<.05$), but no difference between *false photos* and *non-words* (all $p>.05$), (Figure 2).

Language regions:

A whole brain random effects analysis revealed eleven cortical regions that showed greater activation for *false photograph* stories compared to *non-words* (uncorrected, $p<0.001$, $k>10$); these same regions were then identified in individual subjects' data as ROIs using the same threshold: left angular gyrus (identified in 18 out of 18 individuals), left

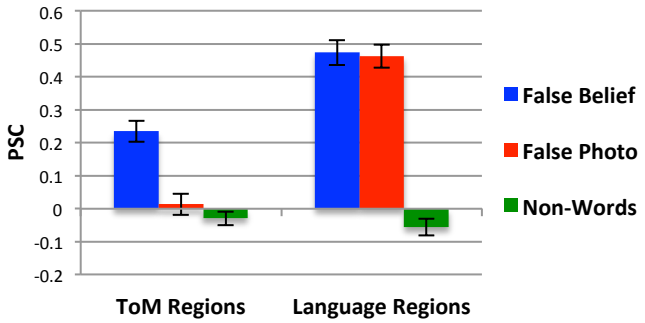


Figure 2: Functional localizer – Cross-Validation results. Average percent signal change in ToM and Language regions for stories about false beliefs, stories about false physical representations, and non-word lists.

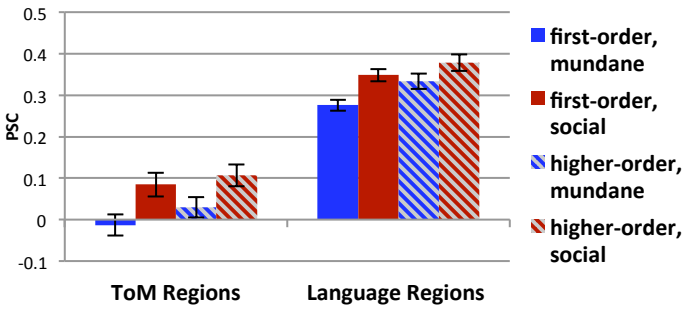
inferior frontal gyrus (18/18), left inferior orbital gyrus (18/18), left medial gyrus (18/18), left superior gyrus (18/18), left anterior lobe (18/18), left middle anterior lobe (18/18), left middle posterior lobe (18/18), left posterior temporal lobe (18/18), right middle anterior lobe (18/18), and right middle posterior lobe (17/18). These regions are those also implicated in a series of previous studies contrasting activation for sentence processing compared to nonsense word processing and to backward speech (Fedorenko et al 2010) (Figure 1b).

As in the ToM regions, the results from the cross-validation were analyzed using pair-wise comparisons of the response to *false belief* stories, *false photograph* stories, and *non-word* lists. Paired-sample two-tail t-tests revealed that all of the individually localized language regions show a significant difference between *false belief* and *non-words* (all p<.05), and between *false photograph* (all p<.05), and *non-words*, but no difference between *false belief* and *false photograph* (all p>.05) (Figure 2).

Experiment 2: Structure and Content

Structure:

Whole brain analysis: A whole brain random effects analysis, contrasting higher-order with first-order beliefs, revealed activity in eight of the eleven previously identified language regions: the left inferior gyrus (IFG) left inferior orbital gyrus (IFGorb), left middle frontal gyrus (MFG), left anterior temporal lobe (ATL), left middle posterior temporal lobe (mPTL), and left and right posterior temporal lobe (PTL) (uncorrected, $p < 0.001$, $k > 10$). None of the voxels in this analysis overlapped with ToM regions. Additionally, we see activation in the area in left IFG located between the left IFG region and left IFG orbital region, as well as in the right counterpart – areas argued to be implicated in response inhibition and working memory (e.g. Aron et al, 2004; Bunge et al 2003, Chikazoe, et al, 2007), the frontal eye fields, dorsolateral prefrontal cortex, and the dorsal anterior cingulate area.



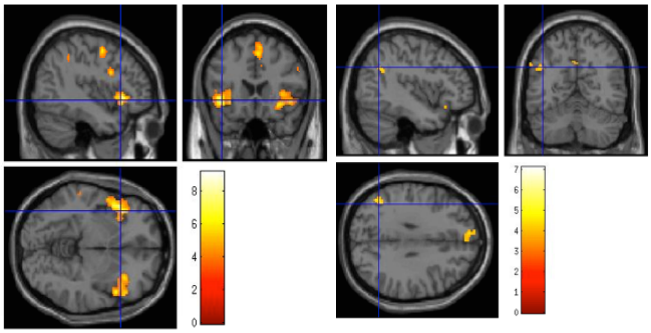


Figure 1 (top): Structure and Content ROI analysis. Average percent signal change in ToM and Language regions for stories varying in their Structure (first-order and second-order) and their Content (mundane and socially relevant)

Figure 2a,b (bottom): Structure and Content whole brain analyses – Embedding (left) and Social Relevance (right). Brain regions in which the bold signal was higher for stories with embedded structure compared to matched stories with first-order structure; and with socially relevant compared to mundane content (N = 18, random effects analysis, $p < 0.001$, uncorrected.)

ROI Percent Signal Change Analysis: Supporting the results of the whole brain analysis, a 2x2 repeated measures ANOVA (Content by Structure) revealed that most of the individually-defined language ROIs were sensitive to the structure of belief stories. *Higher-order* beliefs elicited a significantly higher response than *first-order* beliefs in left IFG, left IFG-orbital, left MFG, left mPTL, and left PTL (all $F>5$, $p<0.01$, $\eta^2>0.25$). The left ATL and right mPTL showed marginal effects in the same direction (both $F>3$, $p<0.1$). One language ROI showed the opposite profile: *first-order* beliefs elicited a higher response than *second-order* ones: the left angular gyrus ($F>5$ $p<0.05$, $\eta^2>0.25$)

By contrast, none of the ToM regions showed a differential response to *second-order* compared to *first-order* belief attributions (all $p >> 0.1$).

These analyses together clearly indicate that the structure of a belief attribution – as encoded by varying the number of embeddings – is not an aspect of belief attribution that affects ToM brain regions. Unlike the ToM regions, most of the language regions show some differentiation between *first-order* and *higher-order* stimuli. Moreover, we see that four of the eleven regions show robustly stronger responses to higher-order beliefs, revealed in both ROI and whole brain analyses: left inferior orbital gyrus, left middle frontal gyrus, left anterior lobe, and left medial posterior lobe. This suggests that this type of structure is an aspect of linguistic stimuli that is represented or processed in a portion of the language processing network, but not the Theory of Mind network.

Content:

Whole brain analysis: A whole brain random effects analysis, contrasting *socially relevant* with *mundane* beliefs, revealed activations that overlapped with four of the five previously identified ToM regions (uncorrected, $p < 0.001$, $k > 10$): left temporo-parietal junction, right superior temporal sulcus, precuneus, and dorsal-medial prefrontal cortex. In addition, there was also activation overlapping with one language region, left inferior frontal gyrus, as well as in the thalamus.

ROI Percent Signal Change Analysis: A 2x2 repeated measures ANOVA (Content by Structure) revealed that all five of the ToM regions identified in the functional localizer showed a main effect of Content, with a significantly higher response to beliefs with *socially-relevant* content, compared to beliefs with *mundane* content. These regions included the right and left TPJ, RSTS, precuneus, and DMPFC (all $F>4$, $p<0.05$, $\eta^2>0.25$).

Nearly all of the language regions also showed a higher response to *socially-relevant* beliefs compared to *mundane*

ones (all $F>4$, $p<0.05$, $\eta^2>0.25$); none of the ROIs showed an interaction between Content and Structure.

General Discussion

We functionally localized both Theory of Mind and higher-level language processing regions in the same individual subjects, and then asked whether changing two distinct aspects of stories describing beliefs modulated the neural response of these two networks. We found that the ToM network is sensitive to the content, but not structural complexity, of stories about beliefs. By contrast, brain regions involved in language processing respond to both the structural complexity and, to some extent, the social relevance of the stories.

In Experiment 1, we presented a new functional localizer for language and Theory of Mind regions, using *false belief* stories, *false photograph* stories, and lists of *non-words*. Following previous work in the Theory of Mind literature, (e.g. Saxe and Kanwisher 2003), we used the contrast of *false belief* stories over *false photograph* stories to localize the Theory of Mind network. In the current localizer, unlike previous ToM localizers, we controlled for a variety of low-level linguistic features, matching the *false belief* and *false photograph* stories on qualities that affect language processing difficulty and linguistic complexity. Despite this, we found all of the classic Theory of Mind regions, including bilateral TPJ, RSTS, PC, and DMPFC, suggesting that these results, and previous results using this type of contrast, are not driven by confounding low-level language features, but rather by the genuine contrast in content – other people's (outdated) mental states versus (outdated) physical representations of the world.

Similarly, following previous work in the neurolinguistics literature (e.g. Fedorenko et al, 2010), we localized regions sensitive to word- and sentence-level processing by doing a *sentences* to *non-words* contrast, using only the *false photograph* stories for the sentences. By matching the *false photograph* stories and *non-words* lists on additional low-level features (bigram frequency and length) and excluding *false belief* stories, we ensured that the localizer was (a) not showing a contrast due simply to increased difficulty in linguistic processing, and (b) not picking out regions that specifically process belief/social information. Using this contrast, we identified eleven cortical regions previously implicated in high-level language processing, including the left IFG, MFG, and SFG, left and right ATL, left PTL, and left angular gyrus.

Finally, by localizing the ToM and language networks in the same participants, we found that the ToM network and language-processing network are both spatially and functionally distinct: ToM regions show a strong BOLD response to *false belief* stimuli, but not to either *false photograph* or *non-word* stimuli; the language regions show an equally strong response to both *false belief* and *false photograph* stories, but not to *non-word* lists.

In Experiment 2, we find that these regions also show different profiles of response to manipulations of the structure and content of stories about belief. We identified two principle dimensions along which descriptions of someone's thoughts can differ, affecting both the mental state itself, and the associated linguistic representation. The first dimension was the content of the thought; here, we manipulated the social relevance of the mental states and events being considered. The second dimension was the structural complexity of the attribution; in this study, we manipulated the number of levels of embedding of the target thought.

We asked whether either (or both) of these dimensions are represented in, and would therefore modulate, the activity of brain regions previously implicated in Theory of Mind and high-level language processing. We found that while both sets of regions were modulated by content, only language regions were affected by the story structure.

Specifically, both whole brain and regions of interest analyses indicated that there was greater activation in Theory of Mind brain regions (temporo-parietal junction, superior temporal sulcus, medial precuneus and dorsal-medial prefrontal cortex) for *socially relevant* mental states than for *mundane* mental states, suggesting that the ToM network appears to be particularly sensitive to the socially relevant stimuli. However, this result must be interpreted with caution, given the similar (if weaker) pattern observed in the language regions in the ROI analysis. The socially-relevant stories in the current experiment were both more arousing and more surprising than the mundane stimuli. Moreover, socially relevant (and in fact, scandalous) information is likely to be more informative, both in making judgments about the belief-holder, and about the world. As a consequence, generally higher responses to the socially relevant stimuli might reflect overall higher arousal or attention, and/or specific representations of the belief content. We are currently doing further work to tease apart the effect of content manipulations in ToM and language regions, and ask what features of "social relevance" might be driving this effect.

In contrast, both whole brain and regions of interest analyses indicated that nearly all of the language regions show sensitivity to the manipulation of structure – left IFG and left orbital IFG, left MFG, left ATL, left and left middle PTL, and right middle PTL. Seeing this effect spread across the extended language network is not very surprising, as the manipulation likely affected a number of different language-related processes, including working memory, syntactic complexity, and semantic complexity. Our results largely converge with Shetreet et al. (2009), who contrasted constructions with no embedding to those will full sentential embedding, and found increased activity in left IFG, bilateral STG (mPTL), bilateral SMG (PTL), left SFG, and left MFG.

In light of this large and general response in the language regions, and the cross-network response to content manipulation, the most interesting result of this paper is the *absence* of sensitivity to belief structure in the ToM regions. Neither whole brain nor regions of interest analyses found evidence that Theory of Mind brain regions' responses are differentially affected by the structural complexity of attributed beliefs.

The fact that language regions, but not ToM regions, show sensitivity to multiple embeddings is particularly surprising given a common assumption that higher-order belief

attributions should invoke more ToM processing – higher-order beliefs require additional meta-representation (the representation of Steve's representation of Jessica's mental state), the representation of more individual thoughts (both Steve and Jessica's), and successful processing of a more complex thought (Steve's) – all things that seem crucially related to ToM representation. Second-order false belief tasks are significantly harder for children than first-order false belief tasks (and are often successfully passed 2-3 years later than first-order false belief tasks) (Apperly et al, 2007; Hollebrandse et al, 2007), have been shown invoke less ToM activation in children (Kobayashi et al., 2007) and been shown to be significantly more difficult for patients with brain damage (Fine, Lumsden, & Blair, 2001) and for individuals with autism or schizophrenia (Baron-Cohen, 1989; Pickup & Frith, 2001).

Thus, given the ToM network's selective response to mental state attributions, we would expect that, if the difference between first-order and higher-order ToM attributions is due to increased complexity of mental state representation, the Theory of Mind network would respond to exactly that sort of complexity. However, while there is clear evidence that an increase in embedding does increase overall processing difficulty and here modulates a variety of language processing regions, we do not see this increased difficulty reflected in the activation of Theory of Mind network.

The fact that an increased number of embeddings does not lead to additional activity in the Theory of Mind regions suggests that the differences between these two types of stimuli, including differences in syntactic complexity, the number of mental states, and meta-representational complexity do not directly drive ToM activity. Rather, we see these differences between first-order and higher-order mental state attribution reflected in increased activation in the areas associated with language and domain general processing. This finding converges with results from patient populations showing that failure to pass second-order false belief tasks may in fact be due to domain-general impairment, rather than diminished theory of mind processing (e.g. Zaitchik, Koff, Brownell, Winner, And Albert, 2006).

This dissociation between mental state embedding and Theory of Mind activation raises questions both cognitively and neurally – are there dimensions of mental state attribution do directly modulate theory of mind processing, do tasks using second-order ToM invoke more theory of mind processing, what parts of belief representation are crucially represented in the ToM network, and what does that suggest about the cognitive architecture of ToM attribution?

Together, our results have started to define and narrow the possible space of Theory of Mind representation. In Experiment 1, we've shown that the neural regions underlying belief attribution are distinct from general language processing regions. In Experiment 2, we show that two dimensions relevant to both belief attribution and language processing affect Theory of Mind and language regions differently. While on one hand, the results suggest that the theory of mind network shows sensitivity to content within the belief attributions, they also clearly show that, despite the obvious link between embedding and mental state attribution, this type of structural complexity does not seem to be part of the representational space of the ToM network.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant 095518, a John Merck Scholars Grant, and a National Science Foundation Graduate Research Fellowship, Grant 0645960.

References

Apperly, I. A., Samson, D., Chiavarino, C., Bickerton, W. -L., & Humphreys, G. W. (2007). Testing the domain-specificity of a theory of mind deficit in brain-injured patients. Cognition, 103(2), 300-321.

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. Journal of Child Psychology and Psychiatry, 30(2), 285-297.

Brunet, E., Sarfati, Y., Hardy-Bayle, M. C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. Neuroimage, 11(2), 157-66.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., Saxe, R. (In Press). fMRI item analysis in a theory of mind task. NeuroimageM.

Fedorenko, E., Hsieh, P.-J., Nieto-Castañon, A., Whitfield-Gabrieli, S. & Kanwisher, N. (2010). A new method for fMRI investigations of language: Defining ROIs functionally in individual subjects. Journal of Neurophysiology, 104, 1177-1194.

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., et al. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. Cognition, 57(2), 109-28.

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.

Gallagher, H.L., Jack, A.I., Roepstorff, A., Frith, C.D., 2002. Imaging the intentional stance in a competitive game. NeuroImage 16, 814–821.

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. J Cogn Neurosci, 19(11), 1803- 14.

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

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539), 2425-30.

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. Soc Neurosci, 1(3-4), 245-58.

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? . Trends Cogn Sci, 10, 59‑63.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". Neuroimage, 19(4), 1835-42.

Shetreet, E., Friedmann, N., & Hadar, U. (2009). An fMRI study of syntactic layers: Sentential and lexical aspects of embedding. NeuroImage, 48(4), 707-716. doi:10.1016/j.neuroimage.2009.07.001

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. Cognition, 116(1), 87-100. Elsevier B.V. doi:10.1016/j.cognition.2010.04.003

Zaitchik, D., Koff, E., Brownell, H., Winner, E., Albert, M. (2006). Inference of beliefs and emotions in patients with Alzheimer. Neuropsychology, 20, 11–20.