MITLibraries | DSpace@MIT

MIT Open Access Articles

Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults

Massachusetts Institute of Technology

Thinking about Seeing: perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults

Jorie Koster-Hale[*,1], **Marina Bedny**[2], and **Rebecca Saxe**[1]

[1]Department of Brain and Cognitive Sciences Massachusetts Institute of Technology Cambridge MA 02139, USA

[2]Department of Psychological and Brain Sciences Johns Hopkins University Baltimore, MD 21218, USA

Abstract

Blind people's inferences about how other people see provide a window into fundamental questions about the human capacity to think about one another's thoughts. By working with blind individuals, we can ask both what kinds of representations people form about others' minds, and how much these representations depend on the observer having had similar mental states themselves. Thinking about others' mental states depends on a specific group of brain regions, including the right temporo-parietal junction (RTPJ). We investigated the representations of others' mental states in these brain regions, using multivoxel pattern analyses (MVPA). We found that, first, in the RTPJ of sighted adults, the pattern of neural response distinguished the source of the mental state (did the protagonist see or hear something?) but not the valence (did the protagonist feel good or bad?). Second, these neural representations were preserved in congenitally blind adults. These results suggest that the temporo-parietal junction contains explicit, abstract representations of features of others' mental states, including the perceptual source. The persistence of these representations in congenitally blind adults, who have no first-person experience with sight, provides evidence that these representations emerge even in the absence of first-person perceptual experiences.

Keywords

blindness; theory of mind; experience; representation; fMRI; multivoxel pattern analysis (MVPA)

[*]**Corresponding author.**jorie@mit.edu**telephone:** +1 857 272 0733 77 Massachusetts Ave 46-4021C Cambridge MA 02210.

Introduction

Imagine a friend tells you that last night, looking out the window onto a dark, rainy street, she saw her boyfriend get into a car with a strange woman, and drive away. Your reaction will depend on many inferences about her thoughts and feelings. You will recognize that she believes her boyfriend is being unfaithful, and feels betrayed. You might also note the source of her belief, and question how clearly she could see at a distance and in the dark. Perhaps she was mistaken about the man getting into the car, or about the driver; maybe the driver was actually her boyfriend's sister. Knowing how she got her information might strongly affect how you reason about her beliefs and experiences -- do you yourself believe that her boyfriend is being unfaithful? How strongly do you think she believes it? What is she likely to do next?

Now imagine that you are congenitally blind. Would your inferences be any different? Clearly, a blind adult would understand the emotional toll of discovering a lover's possible betrayal, but could a blind person make the same inferences about the visual source of the discovery? How much would a blind person understand about the experience of seeing a familiar person and a strange woman, from afar, in the dark?

Blind people's inferences about how other people see provide a window into a fundamental question about the human capacity to think about one another's thoughts: what are the mechanisms used to think about someone else's mind? One possibility is that we think about someone's experience by invoking our own relevant experiences and sensations. In this view, thinking about someone else's experience of seeing requires (among other things) a first-person experience of sight. In contrast, if people use an intuitive "theory" of mind, composed of relationships among abstract mental state concepts, to reason about others' experiences, then experience of sight is not always necessary for reasoning about seeing. In many cases, these views are hard to disentangle; however, they predict different outcomes in blindness. If first-person experience is necessary to understand others' experiences, blind people should have only a fragmentary, limited, or metaphorical understanding of seeing. By asking how blind individuals represent other's experiences of sight, we can place important limits on our theories of mental state inference: to what extent does theory of mind depend on the observer having had similar sensations, experiences, beliefs, and feelings as the target?

The first possibility is that people understand another's mind by trying to replicate it, by imagining themselves in a similar situation, or by re-experiencing a similar past event of their own lives. You would understand your friend's feelings of sadness by imagining your own feelings in response to a lover's betrayal, recreating in your emotional system a version of your friend's experience. Similarly, you would understand your friend's experience of seeing her boyfriend get into the car by recreating the visual scene in your own mind's eye (Gordon, 1989; Stich & Nichols, 1992). Understanding others' minds thus depends on the observer having experienced a *relevantly similar* mental state to the target (Gallese & Goldman, 1998; Goldman, 1989; 2006; Gordon, 1989; Harris, 1992; Nichols, Stich, Leslie, & Klein, 1996; Stich & Nichols, 1992). Simulation-based accounts do not necessarily posit that people can only think about exactly those experiences that they themselves have had;

rather, mental state representation could be a composition of one's existing relevantly similar first-person experiences, composed flexibly to simulate a novel experience. Still, because simulation depends on similar experiences, the extent to which we can simulate the minds of others depends on "the interpersonal sharing of the same kind of neural and cognitive resources. When this sharing is limited (or even missing), people are not fully able (or are not able at all) to map the mental states or processes of others because they do not have suitable mental states or processes to reuse" (Gallese & Sinigaglia, 2011). Because congenitally blind people lack the mental states and processes involved in seeing, their representations of sight are predicted to be limited or unreliable.

Neuroimaging experiments provide evidence that first-person sensorimotor representations are "reused" during observation of others' actions and sensations. Similar brain regions are recruited when experiencing physical pain compared to observing another person experience similar pain (e.g., Botvinick et al., 2005; Immordino-Yang, McColl, Damasio, & Damasio, 2009; Singer et al., 2004); and when experiencing a tactile sensation compared to observing another person being touched in the same way (Blakemore, 2005; Keysers:2004dj Gazzola & Keysers, 2009). More importantly, neural activity during observation depends on the observer's own specific first-person experiences. For example, motor activation in dancers during observation of dance moves is enhanced for specific movements that the observers themselves have frequently executed (Calvo-Merino, Glaser, Grèzes, Passingham, & Haggard, 2005; Calvo-Merino, Grèzes, Glaser, Passingham, & Haggard, 2006). If this type of reuse extends to mental state representation, typical representations of other's experiences of seeing should depend on, or be profoundly affected by, having first seen yourself.

However, many authors (Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1992; Perner, 1993; Saxe, 2005) have suggested an alternative mechanism for understanding other minds: namely, that people have an intuitive theory of other minds. An intuitive theory includes causal relations among abstract concepts (like *belief*s and *desires*), and can be learned from many sources of evidence, not limited to first-person experiences. One source of evidence children could use to build a theory of mind is the testimony of others: verbal labels and descriptions of mental states and experiences, often including mental state verbs like *thin*k and *see* (Harris, 2002b; 1992). Thus a congenitally blind child, growing up in a world full of sighted people, might develop an intuitive theory that includes concepts of vision, to explain everyone else's behavior (e.g. reacting to objects at a distance) and testimony (e.g. saying "I see your toy on the top shelf!"). This intuitive theory would then allow a blind child to predict how a sighted person would act in a given environment, and what that person would be likely to infer based on what she could see.

To test these theories, we investigated how blind people think about sight. Observation and behavioral studies suggest that even young blind children know that other people can see with their eyes, and can understand basic principles of vision: e.g. that objects can be seen from a distance and are invisible in the dark (Bigelow, 1992; Landau & Gleitman, 1985; Peterson, & Webb, 2000). By adulthood, congenitally blind people know the meanings of verbs of sight, including fine-grained distinctions, such as the difference between verbs like *peer, gaze*, and *gawk* (Landau & Gleitman, 1985; Lenci, Baroni, Cazzolli, & Marotta, 2013,

Koster-Hale et al., in prep). Blind adults are thus sensitive to subtle distinctions in how sighted people gather information visually.

This behavioral evidence alone, however, cannot answer the question of whether a blind person uses the same cognitive mechanisms as a sighted person to understand sight. Any surface similarity between a blind and sighted person's verbal descriptions of sight could be the product of compensatory mechanisms in the blind. For example, some authors have suggested that blind people mimic the words used by sighted people, without being able to fully access their meaning (so-called "verbalisms," Rosel, Caballer, Jara, & Oliver, 2005) or integrate them into their conceptual understanding. Thus, a blind person who hears the sentence "I saw my boyfriend getting into the car from the window," may have only a limited or metaphorical understanding of the experience it describes.

This methodological challenge affords an opportunity for cognitive neuroscience: functional neuroimaging can provide an online, unobtrusive measure of ongoing psychological processes and thus offers an alternative strategy to ask if two groups of people are performing a cognitive task using similar or different mechanisms. Here we use neuroimaging to ask whether blind and sighted people rely on similar cognitive mechanisms when they reason about seeing.

Previous studies have shown that thinking about someone else's thoughts (including those based on visual experiences, Bedny et al 2009) increases metabolic activity in a specific group of brain regions often called the 'mentalizing' or theory of mind network. These regions include the medial prefrontal cortex (MPFC), the precuneus (PC), and the bilateral temporal-parietal junction (TPJ), (e.g., Aichhorn et al., 2009; Bedny, Pascual-Leone, & Saxe, 2009; Lombardo, Chakrabarti, Bullmore, Baron-Cohen, & Consortium, 2011; Mason & Just, 2011; Rabin, Gilboa, Stuss, Mar, & Rosenbaum, 2010; Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Spunt, Satpute, & Lieberman, 2010). However, very little is known about what aspects of mental states are represented in these brain regions. Mental experiences have many features, including the content ("boyfriend in stranger's car"), the valence (very sad), and the modality or source of the experience (seen from afar). Neurons, or groups of neurons, within theory of mind brain regions may represent any or all of these features. Thus, two important questions remain open. First, in sighted people, do neurons in any theory of mind region specifically represent the perceptual source of another person's belief? Second, if so, are similar representations present in blind people?

A powerful approach for understanding neural representation is to ask which features of a stimulus are represented by distinct subpopulations of neurons within a region. For example, within middle temporal visual, subpopulations of neurons that respond to visual stimuli moving orientations are spatially organized at a fine spatial scale. Although no individual fMRI voxel (which includes hundred of thousands of neurons) shows an orientation selective response, and therefore overall early visual cortex shows equal average *magnitude* of response to all orientations, it is possible to detect reliably distinct spatial patterns of response across cortex that do distinguish between orientations (Kamitani & Tong, 2006). This technique of looking for reliable *spatial patterns* of fMRI response within a brain region is called multivoxel pattern analysis (MVPA; (Haynes & Rees, 2006; Kriegeskorte &

Bandettini, 2007; Norman, Polyn, Detre, & Haxby, 2006). MVPA can reveal how stimulus categories are processed within a functional region (Peelen et al, 2006; Haynes & Rees, 2006). MVPA has been successfully used to probe the neural basis of many different types of representation, including subjectively perceived directions of motion in ambiguous stimuli, semantic category, emotional affect, and intent when causing harm (Koster-Hale, Saxe, Dungan, & Young, 2013; Mahon & Caramazza, 2010; e.g., Norman et al., 2006; Peelen, Wiggett, & Downing, 2006; Serences & Boynton, 2007).

In the current context, we can ask if source modality (seeing vs hearing) is a relevant feature of that neural representation. If one set of neurons responds more to stories about seeing than hearing, while another (partially distinct) set responds more to stories about hearing than seeing, we have evidence that seeing and hearing are being represented in different ways in that brain region. Measuring the average magnitude of activity in theory of mind brain regions cannot be used to address this question, because the average magnitude may obscure distinct subpopulations of neurons within the regions. We therefore used multivoxel pattern analysis (MVPA) to look for differences in the neural response to stories about hearing and seeing.

In this study, we first asked whether stories about a someone else's hearing and seeing experiences evoke different spatial patterns of response within theory of mind brain regions in sighted individuals. Because very little is known about how theory of mind is represented, cognitively or neurally, this is itself a fundamental question about mental state representation. We then asked whether the same patterns are observed in congenitally blind people. Finding these patterns would provide support for the notion that blind individuals represent mental experiences of seeing in a qualitatively similar manner to sighted individuals. For comparison, we also tested whether theory of mind regions encode a feature of mental states that should not differ between sighted and blind people: the valence (feeling good versus bad).

Methods

Participants

Thirteen sighted members of the larger MIT community participated (8 women; mean age ± SD, 52 years ± 16), all with normal or corrected-to-normal vision. Ten blind individuals participated (5 women; mean age ± SD, 50 years ± 7). Nine blind participants were born blind and one lost sight between the ages of 2 and 3 years. All blind participants reported having at most faint light perception and no pattern discrimination. None were able to perceive shapes, colors, or motion.

One blind participant was ambidextrous and one was left-handed; one sighted participant was left-handed. All were native English speakers and gave written informed consent in accordance with the requirements of the Institutional Review Board at MIT. Participants were compensated $30/hour for their time.

Comparison to Bedny et al (2009)

These data have previously been published, analyzing the magnitude but not the pattern of response in each region, in Bedny et al (2009). Bedny et al (2009) found that theory of mind regions showed equally high responses to stories about hearing and seeing, in both sighted and blind participants. However, measuring the average magnitude across the entire region may obscure distinct subpopulations of neurons within a region. Specifically, the equally high magnitude of response to stories about seeing and hearing observed by Bedny et al (2009) is consistent with three possibilities: (i) neurons in theory of mind brain regions do not distinguish between hearing and seeing in blind or sighted people, (ii) distinct subpopulations of neurons within theory of mind brain regions respond to stories about seeing versus hearing in both sighted and blind participants, reflecting a common representation of the perceptual source of other's mental states, or (iii) distinct subpopulations of neurons respond to seeing versus hearing in sighted but not blind individuals, reflecting different representations in the two groups, depending on their first person experiences. The current analyses allowed us to distinguish between these possibilities.

Note that the current results exclude one blind participant who was included in the previous paper: this participant was born with cataracts and had some light and shape perception during the first ten years of his life. To be conservative, we exclude his data from the current analyses.

Finally, the previous paper included an analysis of reaction time to the behavioral portion of the task (judging how good or bad the protagonist felt at the end of the story), looking at the seeing events, hearing events, and additional control events, all collapsed across valence. In this paper, we are also treating valence as a dimension of interest in the neural data, and so break down the behavioral data by both modality and valence to report reaction time and rating data.

fMRI Protocol and Task

Participants heard 32 stories (each 13 sec long): four in each of eight conditions. Stories in the four conditions of interest described a protagonist's mental experiences, characterized by both a modality-specific source (something seen vs heard) and a modality-independent valence (whether the protagonist felt good or bad). The "seeing" stories described the protagonist coming to believe something as a result of a visual experience, such as seeing a friend's worried face or recognizing someone's handwriting. The "hearing" stories described the protagonist coming to believe something as a result of an auditory experience, such as hearing a friend's worried voice or recognizing someone's footsteps. The stories with negative valence described an event that would make the protagonist feel bad, such as receiving criticism or losing a game; the stories with positive valence described a good event, such as receiving praise or winning a game (**Figure 1**). The remaining four control conditions, which did not describe mental experiences, are not analyzed here (see Bedny et al., 2009 for details).

For each narrative context (e.g. a job interview, dinner with parents-in-law, cleaning a dorm room), we constructed four endings, one in each condition. Thus the stories formed a matched and counterbalanced 2×2 (seeing vs hearing, positive vs negative) design. Individual participants saw each context only once; every context occurred in all conditions, across participants. Word count was matched across conditions (mean length ± SD, 32 words ± 4). Stories were presented in a pseudorandom order, condition order was counterbalanced across runs and subjects, and no condition was immediately repeated. Rest blocks of 11 sec were presented after each story. Eight stories were presented in each 4 min 34 sec run. The total experiment, four runs, lasted 18 min 18 sec.

After each story, participants indicated whether the main character in the story felt *very bad, a little bad, a little good*, or *very good*, using a button press (1-4). Reaction time was measured from the onset of each question.

Theory of Mind and Language Localizer task—Participants also completed a theory of mind and language localizer task. Participants listened to 48 short verbal stories from two conditions: 24 stories requiring inferences about mental state representations (e.g., thoughts, beliefs) and 24 stories requiring inferences about physical representations (e.g., maps, signs, photographs). These conditions were similar in their meta-representational and logical complexity but differ in whether the reader is building a representation of someone else's mental state (See Saxe & Kanwisher, 2003, Dodell-Feder et al. 2011, and Bedny et al., 2009 for further discussion). After each story, participants answered a true/false question about the story. As a control condition, participants listed to 24 blocks of "noise," unintelligible backwards speech created by playing the stories backwards. The task was performed in 6 runs with 12 items per run (4 belief, 4 physical, and 4 backward-speech). Each run was 6 min and 12 sec long. The stimuli for the localizer and both experiments were digitally recorded by a female speaker at a sampling rate of 44,100 to produce 32-bit digital sound files.

Acquisition and Preprocessing

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 T1-weighted structural images in 128 axial slices with 1.33 mm isotropic voxels (TR = 2 ms, TE = 3.39 ms), and functional blood oxygen level dependent (BOLD) data in 30 near-axial slices using 3x3x4 mm voxels (TR=2 s, TE=40 ms, flip angle=90°). To allow for steady state magnetization, the first 4 seconds of each run were excluded.

Data processing and analysis were performed using SPM2 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. The data were motion corrected, realigned, normalized onto a common brain space (Montreal Neurological Institute (MNI) template), spatially smoothed using a Gaussian filter (full-width half-maximum 5 mm kernel) and subjected to a high-pass filter (128 Hz).

Motion and Artifact Analysis

To estimate motion and data quality, we used three measures for each participant: mean translation was defined as the average absolute translation for each TR across the x, y, and z plane; mean rotation was defined as the average absolute rotation per TR across yaw, pitch, and roll; and number of outliers per run was defined as the average number of time points per run in which (a) TR-to-TR head movement exceeded 2mm of translation, or (b) global mean signal deviated by more than three standard deviations of the mean. One blind participant moved excessively during the scan (mean translation of 1.2 mm, mean rotation of 1.4 degrees, and mean outliers per run = 20.7, compared to other blind participants' means of 0.33 mm of rotation, 0.35 degrees of rotation, and 2.5 outliers), so his results were dropped from the analyses.

fMRI Analysis

All fMRI data 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. A slow event-related design was used. An event was defined as a single story, the event onset was defined by the onset of the story sound file, and offset as the end of the story.

Functional Localizer—Individual subject ROIs In each participant, functional regions of interest (ROIs) were defined in right and left temporo-parietal junction (RTPJ, LTPJ), medial precuneus (PC), dorsal medial prefrontal cortex (DMPFC[3]), and ventral medial prefrontal cortex (VMPFC[3]). Each subject's contrast image (Belief > Photo) was masked with each of the regions' likely locations, using probabilistic maps created from a separate dataset.The peak voxel that occurred in a cluster of 10 or more voxels significant at p<0.001 was selected. All voxels within a 9mm radius of the peak voxel, individually significant at p<0.001, were defined as the ROI.

Within-ROI Pattern Analysis—Split-half correlation-based MVPA asks whether we can find evidence that the neural pattern in a region is sensitive to a category-level distinction. Specifically, we ask whether the neural patterns generated by items within a condition are more similar to each other ("within-condition correlation") than to the neural patterns generated by items in the other conditions ("across-condition correlation"). If we find that within-condition correlations are reliably higher than across-condition correlations, we can conclude that there are reliable neural pattern generated by difference items within a condition, and that these neural patterns are distinct from one condition to another. Together, this suggests that the region is sensitive to the category distinction -- items within a category are coded in a similar way, with distinguishable codes for different categories.

Here, we conducted within-ROI pattern analyses, independently testing for information about belief source and valence in the regions identified in the independent functional

---

[3]Note that in Bedny et al., 2009, DMPFC was called SOFC, and VMPFC was called OFC.

localizer. To compare seeing and hearing beliefs, we collapsed across good and bad valence; to compare good and bad valence, we collapsed across seeing and hearing.

Following Haxby et. al. (2001), each participant's data were divided into even and odd runs ('partitions') and then the mean response (beta value) of every voxel in the ROI was calculated for each condition. Because each participant read 8 stories about hearing, seeing, feeling good, and feeling bad, each partition contained the average response to 4 individual stories. The "pattern" of response was the vector of beta values across voxels within the participants individual ROI. To determine the within-condition correlation, the pattern in one (e.g. even) partition was correlated with the pattern for the same condition in the opposite (e.g. odd) partition; to determine the across-condition correlations the pattern was compared to the opposite condition, across partitions (**Figure 2**).

For each condition pair (e.g. seeing vs. hearing) in each individual, an index of classification was calculated as the within-condition correlation (e.g. the correlation of one half of the seeing stories to the other half of seeing stories, averaged with the correlation of one half of the hearing to the other half of hearing stories) minus the across-condition correlation (e.g. the correlation of seeing stories compared to hearing stories). To allow for direct comparison of correlation coefficients, we transformed all r values using Fisher's Z transform. A region successfully classified a category of stimuli if, across individuals, the within-condition correlation was higher than the across-condition correlation, using a Student's T complementary cumulative distribution function.

This procedure implements a simple linear decoder. Linear decoding, while in principle less flexible and less powerful than non-linear decoding, is preferable both theoretically and empirically. A non-linear classifier can decode nearly any arbitrary feature contained implicitly within an ROI, reflecting properties of the pattern analysis algorithm rather than the brain, which makes successful classification largely uninformative (Cox & Savoy, 2003; DiCarlo & Cox, 2007; Goris & Op de Beeck, 2009; Kamitani & Tong, 2005; Norman et al., 2006). Moreover, linear codes have been argued to be a more neurally plausible way of making information available to the next layer of neurons (Bialek, Rieke, Van Steveninck, & Warland, 1991; Butts et al., 2007; DiCarlo & Cox, 2007; Naselaris, Kay, Nishimoto, & Gallant, 2011; Rolls & Treves, 2011).

Whole Brain Pattern Analysis (Searchlight)—In the searchlight analysis, rather than using a predefined ROI, a Gaussian kernel (14mm FWHM, corresponding approximately to the observed size of the functional ROIs) was moved iteratively across the brain. Using the same logic as the ROI-based MVPA, we computed the spatial correlation, in each kernel, of the neural response (i.e. betas) within conditions and across conditions. We then transformed the correlations using Fisher's Z, and subtracted the across-condition from the within-condition correlation to create an index of classification. Thus, for each voxel, we obtained an index of how well the spatial pattern of response in the local region (i.e. the area centered on that voxel) can distinguish between the two conditions. The use of a Gaussian kernel smoothly de-emphasizes the influence of voxels at increasing distances from the reference voxel (Fedorenko, Nieto-Castañón, & Kanwisher, 2012). We created whole brain maps of the index of classification for each subject. These individual-subject correlation maps were

then subjected to a second-level analysis using a one-sample t-test (thresholded at p<0.001, voxelwise, uncorrected).

Results

Behavioral

Sighted participants—We performed a 2×2 ANOVA, crossing valence (good/bad) and modality (seeing/hearing), on both goodness ratings and reaction time data. Using a 1 (very bad) to 4 (very good) scale, sighted participants rated protagonists who experienced positive events as feeling better than the protagonists experiencing negative events, with no effect of modality and no interaction (hearing-bad: 1.84±0.14, hearing-good: 3.43±0.14, seeing-bad: 1.88±0.12, seeing-good: 3.65±0.11; main effect of valence: $F(1,12)=223$, $p<0.001$, partial $\eta^2=0.79$; modality: $F(1,12)=1.2$ $p=0.3$, partial $\eta^2=0.02$; modality by valence interaction: $F(1,12)=0.4$, $p=0.5$, partial $\eta^2=0.01$).

Sighted participants showed small, but significant effects of condition on reaction time (measured from onset of the question), with a marginal main effect of modality and a small but significant interaction (hearing-bad: 5.63±0.23, hearing-good: 5.2±0.21, seeing-bad: 5.03±0.3, seeing-good: 5.28±0.23; main effect of modality: $F(1,12)=4.4$, $p=0.06$, partial $\eta^2=0.02$; valence: $F(1,12)=0.3$, $p=0.5$, partial $\eta^2<0.01$; modality by valence interaction: $F(1,12)=6.8$, $p=0.002$, partial $\eta^2=0.04$). Post-hoc t-tests reveal that sighted participants responded more slowly to stories about negative experiences based on hearing than based on seeing ($t(12)=3.3$, $p=0.007$); there was no effect of modality on responses to positive events ($t(12)=0.47$, $p=0.65$).

Blind participants—Congenitally blind participants also rated protagonists in positive stories as feeling significantly better than those in negative stories, with no effect of modality and no interaction (hearing-bad: 1.81±0.15, hearing-good: 3.51±0.13, seeing-bad: 1.95±0.25, seeing-good: 3.43±0.11; main effect of valence: $F(1,8)=72$, $p<0.001$, partial $\eta^2=0.72$; modality: $F(1,8)=0.02$, $p=0.9$, partial $\eta^2<0.01$; modality by valence interaction: $F(1,8)=0.5$, $p=0.5$, partial $\eta^2<0.02$).

In reaction time of the blind adults, there were no significant effects of modality or valence, and no interaction (hearing-bad: 5.51±0.35, hearing-good: 5.02±0.41, seeing-bad: 5.47±0.33, seeing-good: 5.64±0.3; all $F<3.1$, all $p>0.1$).

Across Groups—We found no differences in the ratings across groups. We performed a 2×2×2 repeated measures ANOVA crossing valence and modality as within-subjects factors with group (blind/sighted) as a between-subjects factor. All participants rated the protagonist as feeling worse in the negative valence stories ($F(1,20)=264$, $p<0.001$, partial $\eta^2=0.76$), with no main effect of modality or group, and no interactions (all $F<1$, $p>0.3$).

In the reaction time data, there were no main effects of modality, valence, or group (all $F<1$, $p>0.4$). We found a small but significant modality by valence interaction ($F(1,20)=7.0$, $p=0.02$, partial $\eta^2=0.03$): participants responded more slowly to stories about negative

experiences based on hearing than seeing; and more slowly to stories about positive experiences based on seeing than hearing.

There was also a small group by modality interaction ($F(1,20)=7.4$, $p=0.01$, partial $\eta^2=0.02$): Sighted adults respond faster than blind adults to stories about seeing; blind adults respond faster than sighted adults to stories about hearing. Post-hoc t-tests comparing groups within modality (e.g. RTs of blind vs sighted for seeing stories) revealed no differences in reaction time between groups in either modality (all $t<1.1$, all $p>0.2$).

fMRI results

Motion and Artifact Analysis Results—Sighted participants and blind participants showed no difference in the mean translation per run (sighted mean + sd = 0.22mm ± 0.03, blind = 0.23mm ± 0.04), $t(20)=0.13$, $p=0.9$), mean rotation per run (sighted = 0.26 degrees ± 0.04, blind = 0.24 ± 0.04, $t(20)=0.37$, $p=0.71$, or the mean number of outliers per run (sighted = 0.98±0.46, blind = 0.53±0.35, $t(20)=0.72$, $p=0.48$) Together, these data suggest the groups were well matched in motion and scanner noise.

Localizer—Replicating many studies using a similar functional localizer task (e.g. Saxe & Kanwisher, 2003), we localized five theory of mind brain regions showing greater activation for false belief stories compared to false photograph stories in the majority of participants (uncorrected, $p<0.001$, $k>10$): **Sighted participants:** RTPJ 13/13 participants, LTPJ, 12/13, PC 13/13, DMPFC 11/13, VMPFC 11/13; **Blind participants:** RTPJ 8/9 participants, LTPJ 9/9, PC 9/9 and DMPFC 9/9, VMPFC 9/9 (**Figure 3A**). As reported in (Bedny et al., 2009), sighted and blind participants did not differ in the activation or anatomical loci of any active regions (all $p>0.1$), nor did blind participants show more variability in spatial location or size of ROIs.

Within ROI Pattern Analysis

Sighted participants

<u>**Source (seeing vs. hearing):**</u> Multi-voxel pattern analyses revealed reliably distinct patterns of neural activity for stories about seeing versus hearing in the RTPJ and LTPJ, but not PC, DMPFC, or VMPFC, of sighted adults. Note that correlations are Fisher Z transformed to allow statistical comparisons with parametric tests. Across partitions of the data, the pattern generated by stories in one category (seeing or hearing) was more correlated with the pattern for the same category than with the pattern for the opposite category, in the RTPJ (within condition correlation ± standard error, $z=1.1±0.2$, across condition correlation, $z=0.95±0.2$, $t(12)=2.0$, $p=0.03$) and LTPJ (within = 0.82±0.2, across = 0.64±0.2, $t(11)=2.0$, $p=0.04$), but not in the PC (within = 0.86±0.2, across = 0.78±0.2, $t(12)=0.93$, $p=0.19$), DMPFC (within=0.7±0.1, across=0.5±0.2, $t(10)=1.2$, $p=0.12$) or VMPFC (within = 0.58±0.2, across = 0.47±0.2, $t(10)=0.75$, $p=0.23$, **Figure 3B, Table 1**).

To test whether the difference between ROIs itself was significant, we conducted a 1x5 repeated measures ANOVA. Because baseline differences in correlations across ROIs are hard to interpret (higher overall correlation in one region compared to another could be due to many factors, including the distance of a region from the coils, amount of vascularization,

or region size; see Smith, Kosillo, & Williams, 2011), we used difference scores (within condition correlation - across condition correlation) as the dependent variable. We found a main effect of ROI (F(4,28)=3.03, p=0.03, partial η^2=0.27), suggesting that the regions contain varying amounts of information about source modality in their neural pattern.

Valence (good vs. bad): In no region did the pattern of response distinguish between good and bad valence (all correlation differences t<0.2, all p>0.1, **Figure 3B**).

Blind participants

Source (seeing vs. hearing): The pattern generated by stories within a condition were more correlated with other stories in the same condition compared to stories in the opposite condition in the RTPJ (within condition correlation ± standard error, z=1.1± 0.2, across condition correlation z=0.93±0.3, t(7)=2.0, p=0.04), but not in the LTPJ (within=1.3±0.3, across=1.3±.03, t(8)=0.2 p=0.4), PC (within=0.71±0.2, across=0.59±0.2, t(8)=1.3, p=0.12), DMPFC (within=0.83±0.2, across=0.73±0.1, t(8)=1.4, p=0.11) or VMPFC (within=0.56±0.2, across=0.34±0.2, t(8)=1.4, p=0.11, **Figure 3C, Table 1**). This difference between ROIs in discrimination was significant (F(4,28) = 4.0, p=0.01, partial η^2 =0.36).

Valence (good vs. bad): As in sighted adults, the pattern of response in congenially blind adults did not distinguish between good and bad valence in any theory of mind region (all correlation differences t<0.2, all p>0.1, **Figure 3C**).

Across Groups

Source (seeing vs. hearing): Comparing across groups, we looked for three things: (a) a main effect of discrimination, driven by differences between the within-condition and across-condition correlations in that region: evidence of reliable discrimination between conditions; (b) a main effect of group, which indicates that one group has overall higher correlations, due to higher overall inter-trial correlations independent of condition (and thus not suggestive of interpretable group differences), and (c) an interaction between discrimination and group, such that one group shows a larger difference between the within- and across-condition correlations: evidence that one group shows more sensitivity to condition differences than the other.

Overall, we found that blind and sighted adults showed very similar neural patterns, with no evidence that either group was more sensitive to the distinction of seeing versus hearing. We found evidence of distinct neural patterns for seeing and hearing in both blind and sighted adults in the TPJ and no other regions. Specifically, blind and sighted participants show equally robust neural discrimination of seeing versus hearing in the RTPJ, with a main effect of discrimination (F(1,19) = 7.8, p=0.01, partial η^2 = 0.3), no effect of group (F(1,19)=0.001, p=0.9), and, critically, no interaction (F(1,19)=0.02, p=0.89). There were no significant effects of group, discrimination, or their interaction in any other ROI. LTPJ showed a trend towards both a main effect of group (overall higher inter-trial correlations in blind participants, F(1,19)=3.2, p=0.09) and a main effect of discrimination (F(1,19)=3.6, p=0.07), with no interaction. Precuneus, DMPFC, and VMPFC showed no effects (PC: group: F(1,20)=0.5, p=0.5, discrimination: F(1,20)=2.3, p=0.14, interaction: F(1,20)=0.1,

p=0.7; DMPFC: group: F(1,18)=0.9, p=0.4, discrimination: F(1,18)=2.7, p=0.12, interaction: F(1,18)=0.3, p=0.6; VMPFC: group: F(1,18)=0.1, p=0.8, discrimination: F(1,18)=2.1, p=0.16, interaction: F(1,18)=0.3, p=0.6)

Valence (good vs. bad): We found no evidence of neural code distinguishing good valence from bad in any brain region of either group. Comparing patterns for valence, we found a main effect of group in the LTPJ (F(1,18)=5.1, p=0.04, partial η^2=0.22), due to higher overall inter-trial correlations in the blind, independent of condition (and thus not suggestive of sensitivity to differences in valence). There were no significant effects of group, discrimination, or their interaction in any other ROI.

Whole Brain Pattern Analysis

Source (seeing vs. hearing)—The results of the searchlight converge with the ROI analyses, suggesting a representation of seeing and hearing in the RTPJ of both groups. Because of the similarity across groups in the ROI analyses, we combined the data from blind and sighted participants for increased power. Converging with the results of the ROI analyses, the whole brain analysis revealed that only the RTPJ distinguished between seeing versus hearing beliefs (n = 22, peak voxel at [60, −48, 12], p<0.001, uncorrected peak T = 3.6).

Second, a two-sample T-test across groups revealed a significant group difference in the left dorsolateral PFC (BA45/46, peak voxel at [−58, 30, 8], p<0.001 uncorrected, peak T = 3.7). In this DLPFC region, sighted participants showed a greater difference between within and across condition correlations than blind participants. No regions showed stronger decoding in blind participants relative to sighted ones.

Valence (good vs. bad)—Combining across groups for power, we find that only the anterior cerebellum distinguishes between good and bad emotional valence (n = 22, peak voxel at [12, −42, −32], p<0.001 uncorrected, peak T = 4.2).

General Discussion

MVPA reveals features of mental state representations

For neuroimaging research to have cognitive implications, a key challenge is to go beyond where a cognitive function occurs in the brain, to provide a window into neural representations and computations. Dozens of neuroimaging studies suggest a hypothesis for where in the brain key aspects of theory of mind are processed (e.g., Aichhorn et al., 2009; Bedny, Pascual-Leone, & Saxe, 2009; Lombardo, Chakrabarti, Bullmore, Baron-Cohen, & Consortium, 2011; Mason & Just, 2011; Rabin, Gilboa, Stuss, Mar, & Rosenbaum, 2010; Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Spunt, Satpute, & Lieberman, 2010). The right and left TPJ, the PC, and MPFC show increased hemodynamic responses to stimuli that require participants to think about other people's mental states, and many studies have investigated the selectivity and domain specificity of these brain regions for theory of mind. These brain regions therefore provide a prime opportunity to probe the neural computations of mental state representation: What features of people's beliefs and intentions are represented, or made explicit, in these brain regions?

A powerful, if simplifying, assumption about neural representation is that a feature is explicitly represented by a population of neurons if that feature can be decoded linearly from the population's response. Different subpopulations of neurons within a region may contribute to a common task by representing different features or aspects of the stimulus or task. A well-studied example is object recognition in the ventral pathway of macaques. Low-level features of the stimulus, like retinotopic position and motion energy, can be linearly decoded from the population response in early visual cortex, whereas higher-level properties, like object identity (invariant to size and position), can be linearly decoded from the population response in IT, a region further along the processing stream (DiCarlo et al. 2012; Kamitani & Tong 2005).

We can take advantage of this property of neural populations to identify features of human neural representations using fMRI. When the subpopulations of neurons that respond to different stimulus features are at least partially organized into spatial clusters or maps over cortex (Dehaene & Cohen, 2007; Formisano et al., 2003), those features may be detectable in reliable spatial patterns of activity measurable with fMRI (Haynes & Rees, 2006; Kriegeskorte & Bandettini, 2007; Norman et al., 2006). Multi-voxel pattern analyses therefore offer an exciting opportunity for studying the representations underlying human theory of mind. In spite of many studies on the neural basis of theory of mind, very little is known about which features or dimensions of mental states are represented by the neural populations in the candidate brain regions. Here we identified two dimensions of others' mental states that are common and salient: the source modality of the other person's experience, and the emotional valence. To ask whether either of these dimensions are explicitly represented in these brain regions, we then tested whether either of these dimensions could be linearly decoded from the spatial pattern of the response in any theory of mind brain region.

We found that both the right and left TPJ of sighted people showed distinct spatial patterns of responses to stories that described seeing versus hearing. Two independent groups of stories, similar only in the source modality of the character's beliefs, elicited reliably similar spatial patterns of response in the TPJ, in spite of the wide heterogeneity of the stories in many other respects (e.g. the physical environment, the type of event, the character's name, age, gender, social status, etc). Furthermore, we replicated these results in the right TPJ in an independent group of congenitally blind adults. We therefore suggest that spatially distinguishable neural populations in the TPJ explicitly represent the perceptual source of another person's mental states. Together with prior evidence, our findings suggest that these representations develop in the absence of first person experience.

Knowledge source and theory of mind

Why might the TPJ code the distinction between seeing and hearing mental states described in stories? One possibility is that the patterns of activation we observed were driven simply by a response to the mental state attitude verb used in the story (i.e. "sees" vs "hears"). However, we consider this interpretation unlikely, for three reasons. First, mental state verbs on their own elicit little to no response in the TPJ, both relative to rest and relative to non-mental verbs (e.g. "to think" vs "to run" or "to rust," Bedny, McGill, & Thompson-Schill,

2008). Robust responses are best elicited in TPJ by full propositional attitudes (e.g., a person's mental attitude to a state), ideally embedded in an ongoing narrative. Second, in previous research we directly manipulated the mental state verb in a sentence and did not find distinct patterns of response in TPJ (e.g. "believe" vs "hope", Paunov and Saxe, personal communication). Third, in an other experiment, we found distinct patterns of response in TPJ for distinct mental states that were described using the same verb (i.e. intentional vs accidental harm, both described by "believed", Koster-Hale et al 2013). In all, we suggest that the patterns of activity observed here are most likely responses to features of the character's mental state, rather than to specific verbs in the stories. Still, future research should test this possibility by independently manipulating the source of the mental state and the verb used to describe it, as well as by looking at the pattern evoked by these verbs outside of a mental state context.

Based on the current findings, we hypothesize that the TPJ contains an explicit representation of the perceptual source of mental states. Why might theory of mind brain regions code this information? Considering the source of others' knowledge is central to inferences about other people's minds. Knowing how another person got their information can influence the inferences we make about what they know, how well they know it, and whether we should believe it ourselves. For example, we are more likely to trust another person's testimony if they acquired their knowledge by direct visual observation (an "eye-witness") than through gossip or hearsay (Bull Kovera, Park, & Penrod, 1992; Peter Miene, Bordiga, & Park, 1993; Peter Miene, Park, & Bordiga, 1992); eyewitness but not hearsay testimony is admitted as evidence in courts of law (Park, 1987). Similarly, recalling how one acquired one's own information plays an important role in evaluating and justifying a belief, and deciding how readily it should be discarded (O'neill & Gopnik, 1991).

Recognizing sources of knowledge is so cognitively salient that some languages explicitly mark source syntactically. In languages with evidential marking, such as Turkish, Bulgarian, Tibetan and Quechua, utterances are syntactically marked to indicate how the information was acquired (Aikhenvald, 2004; J. G. de Villiers, Garfield, Gernet-Girard, Roeper, & Speas, 2009; Faller, 2002; Fitneva, 2001; Johanson & Utas, 2000; Smirnova, 2012). Some languages mark direct versus indirect experience, or hearsay versus everything else; others mark whether the information was gained from inference, hearsay, or sensory experience. Languages such as Tucano and Fasu encode aspects of sensory modality, including distinctions between visual versus non-visual sensory information; one language, Kashaya, has a separate marker for auditory evidence (Aikhenvald, 2004). Acquisition of linguistic evidential marking by children appears to depend on the development of theory of mind: during acquisition of Turkish, for example, children's earlier performance on explicit theory of mind tasks concerning sources of knowledge predict their later comprehension of linguistic evidentials, but not vice versa (Papafragou & Li, 2001).

In sum, the source of others' knowledge is both behaviorally relevant and cognitively salient; the current results suggest that it is also explicitly encoded by distinct neural populations within the TPJ. An open question for future work is whether patterns in the TPJ can distinguish only the modality or also other aspects of another person's source of knowledge. For example, sources of knowledge can be direct (perceptual observation),

inferential (by induction from clues or patterns), or from hearsay (from other people's report). Direct perceptual sources may be more or less reliable, depending on the situation (e.g. observed from nearby or at a distance, based on a quick glance or a long stare) and the observer (e.g. an expert, an amateur, or a child). Future research should investigate the neural representation of these other features of knowledge source, and the relationship between these features and perceptual source.

The valence of others' mental states

In contrast to perceptual source, we failed to decode the valence of another person's mental experience from the pattern of activity in any theory of mind region. This is especially noteworthy, considering that the behavioral task -- judging how good or bad someone felt at the end of the story -- specifically drew attention towards valence, while ignoring source information. However, a prior study successfully decoded positive versus negative valence of others' emotions, expressed in emotional body movements, facial expressions, and tones of voice, from patterns of activity in MPFC (Peelen, Atkinson, & Vuilleumier, 2010). One possibility is that emotional experience is more effectively conveyed in non-verbal stimuli. However, null results from MVPA should be interpreted with caution. The neural sub-populations that respond to different types of valence could be intermingled such that we would be unable to detect the distinct neural populations at the spatial scale of fMRI voxels; the distinction would only be revealed by techniques with higher spatial resolution.

Blind adults represent knowledge source

We found similar patterns of neural activity in blind and sighted adults. Most notably, we found that other people's visual versus auditory experiences are encoded distinctly by both sighted and blind adults. Blind individuals specifically lack first person experience of visual aspects of mental states, and thus allow us to test the role of first person, sensory experience in the development of theory of mind representation: what does it take to be able to think about someone else's experience?

One set of theories of how we understand others' mental experiences posits that we vividly simulate, or imagine, having the same experience ourselves. For example, Gallese & Sinigaglia (2011) describe the process of simulation as follows: "People reuse their own mental states or processes in functionally attributing them to others, where the extent and reliability of such reuse and functional attribution depend on the simulator's bodily resources and their being shared with the target's bodily resources." (p. 518). That is, we understand other people's mental states using a template of our own first-person experiences. This view is difficult to reconcile with evidence that blind people use similar neural mechanisms to reason about seeing as do sighted people. In the initial analyses of these data, Bedny et al. (2009) showed that blind adults can represent mental states acquired by vision, without additional costs or compensatory mechanisms. The present analyses further show that in these same blind and sighted adults, not only are the same brain regions recruited to think about others' mental states based on seeing and hearing, but these regions represent the difference between visual and auditory perceptual sources. These results suggest the representation of perceptual source in the TPJ is not dependent on first-person experience.

Importantly, however, our data do not show that people *never* use first-person experience to make inferences about other people's minds and goals. Both in the motor and sensory domains, people do seem to use sensory-motor information to understand other's actions and experiences. For example, interference between other people's sensory-motor experience and our own suggests that we use partially shared representations for action observation and action execution. Observing someone else's actions can interfere with executing actions yourself (Kilner, Paulignan, & Blakemore, 2003), and observing what another person can see interferes with the ability to report on what you yourself are seeing (Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010). Similarly, personal experience specifically helps children learn about others' visual experiences: infants who experienced a transparent blindfold follow the gaze of a blindfolded adult, but infants who experienced an opaque blindfold do not (Meltzoff, 2004; 2007; Meltzoff & Brooks, 2007).

In all, understanding of other minds is likely guided by both first-person experience, which provides rich and detailed input about the character of specific experiences, and an intuitive theory of mind, which allows individuals to form representations of experiences they have not had, and indeed could never have, such as blind adults' representations of experiences of sight. In some cases, such as action prediction, first person sensory machinery may play a direct role in representing information about other people; in other cases, first person experience provides a rich source of data about how people feel and act, playing an important role in building causal theories of other's minds. The best "theory of mind" will both be able to deal with abstract generalizations and make use of data from first-person experience (Apperly, 2008; Keysers & Gazzola, 2007; Kilner & Frith, 2007).

Learning about sight

The present findings, along with prior evidence, suggest that blind people can have impressively rich knowledge about sight (Landau & Gleitman, 1985; Marmor, n.d.; Shepard & Cooper, 1992). How do blind individuals acquire this extensive knowledge about seeing? We propose that one source of information is others' testimony. Blind children live immersed in a world of sighted people, who talk constantly about experiences of seeing. In general, children acquire much of their knowledge of the invisible causal structure of the world through testimony (Harris, 2002a; 2002b; Harris & Koenig, 2006; Koenig, Clément, & Harris, 2004). Language and testimony are a particularly powerful tool for learning about the invisible contents of other minds (Bedny & Saxe, 2012; Robinson & Whitcombe, 2003; Robinson, Haigh, & Nurmsoo, 2008; Urwin, 1983); the absence of linguistic access to other minds can significantly delay theory of mind development, for example in deaf children born to non-signing parents (Figueras-Costa & Harris, 2001; Meristo et al., 2007; de Villiers, 2005; Moeller & Schick, 2006; Peterson & Wellman, 2010; Peterson & Siegal, 1999; Pyers & Senghas, 2009; Schick, de Villiers, de Villiers, & Hoffmeister, 2007; Woolfe & Want, 2002). In contrast, conversational access seems to give blind children a detailed representation of visual mental states, and the inferences they afford.

On the other hand, these experiments have just begun to probe sight understanding in blind individuals. In the current stimuli (and unlike our initial example of the woman at the window) neither the content nor the reliability of the mental states depended on the

perceptual modality. The belief content (e.g. "his team lost") was matched across perceptual sources, and always reliably inferred from the perceptual evidence. Thus, successfully encoding the story did not depend on specific knowledge about sight, e.g., how difficult it is to recognize a face in the dark. It is possible that blind people would show different behavior and neural processes for such inferences. Additionally, participants' task (judging how good or bad the protagonist would feel) depended on encoding the content but not the source of beliefs. Despite this, participants spontaneously encoded the perceptual source of the belief described in the story. However, as a consequence, the current results cannot determine whether blind and sighted people would show similar neural patterns (or behavioral performance) if specific details about the perceptual source were more relevant in the task. Future research should investigate whether blind and sighted adults have similar knowledge and representations of the details of sight, such as how knowledge based on vision will vary with distance, occlusion, and darkness, and whether the neural patterns observed here encode similar information about seeing and hearing. Such specific details may show greater influences of first person experiences.

The importance of testimony to the understanding of other minds is highlighted by the contrast between how much blind people know about sight and how little sighted people know about blindness. Blind individuals can never directly experience vision, but they are constantly exposed to testimony about seeing. By contrast, sighted people do experience the perceptual aspects of blindness (when in the dark or with their eyes closed). Nevertheless, the average sighted person has little understanding of blindness. Most sighted people do not know how blind individuals navigate their environment or perform daily tasks such as dressing and eating. This lack of knowledge can lead to incorrect and sometimes problematic inferences about blind people's incapability to work, learn, and parent. Given how much blind people know about seeing, such misunderstandings are unlikely to result from an in principle incapacity to understand experiences different from our own. While blind people are surrounded by sighted people, most sighted individuals have never met anyone who is blind. The ignorance of sighted people is thus another dramatic example of the importance of testimony for learning about other minds.

Finally, while the data here suggest a common endpoint in the development of an adult theory of mind, the processes by which blind and sighted children acquire this theory may be different. Although blind adults appear to have typical theory of mind, blind children are delayed on a variety of theory of mind tasks, including both tasks that rely directly on inferences about vision, such as perspective taking tasks (Bigelow, 1991; 1992; Millar, 1976), and also tasks that do not depend directly on visual perspective, including auditory versions of the false belief task (McAlpine & Moore, 1995; Minter, & Hobson, 1998; Peterson et al., 2000). Using a battery of theory of mind tasks designed specifically for blind children, Brambring & Asbrock (2010) concluded that blind children are delayed an average of 1-2 years, relative to blindfolded sighted control children, though these children catch up in adolescence (see Peterson & Wellman, 2010 for a similar argument about deaf children). Thus an open question is whether the neural mechanisms for theory of mind, which are similar in blind and sighted adults, develop differently in blind and sighted children.

Conclusion

In summary, we find that (i) theory of mind brain regions (specifically, the TPJ) encode perceptual source and sensory modality of others' mental states, and (ii) these representations are preserved in the RTPJ of congenitally blind adults. Considerable neuroimaging work on theory of mind suggests that the RTPJ plays a role in thinking about others' thoughts; we find that one aspect of this role is to make explicit, in the population response of its neurons, features of beliefs -- in this case, the perceptual source of the belief. The persistence of these representations in congenitally blind adults, provides evidence that theory of mind brain regions come to encode these aspects of mental states even in the absence of first-person experience.

Acknowledgments

References

Aichhorn MM, Perner J, Weiss B, Kronbichler M, Staffen W, Ladurner G. Temporo-parietal junction activity in theory-of-mind tasks: falseness, beliefs, or attention. Journal of Cognitive Neuroscience. 2009; 21(6):1179–1192. doi:10.1162/jocn.2009.21082. [PubMed: 18702587]

Aikhenvald, AY. Evidentiality. Oxford University Press; 2004.

Apperly IA. Beyond Simulation-Theory and Theory-Theory: why social cognitive neuroscience should use its own concepts to study "theory of mind". Cognition. 2008; 107(1):266–283. doi:10.1016/j.cognition.2007.07.019. [PubMed: 17868666]

Bedny M, Saxe R. Insights into the origins of knowledge from the cognitive neuroscience of blindness. Cognitive Neuropsychology. 2012; 29(1-2):56–84. doi:10.1080/02643294.2012.713342. [PubMed: 23017086]

Bedny M, McGill M, Thompson-Schill SL. Semantic Adaptation and Competition during Word Comprehension. Cerebral Cortex. 2008; 18(11):2574–2585. doi:10.1093/cercor/bhn018. [PubMed: 18308708]

Bedny M, Pascual-Leone A, Saxe RR. Growing up blind does not change the neural bases of Theory of Mind. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(27):11312–11317. doi:10.1073/pnas.0900010106. [PubMed: 19553210]

Bialek W, Rieke F, Van Steveninck RRR, Warland D. Reading a neural code. Science. 1991; 252(5014):1854–1857. [PubMed: 2063199]

Bigelow A. Hiding in blind and sighted children. Development and Psychopathology. 1991; (3):301–310.

Bigelow A. The development of blind children's ability to predict what another sees. 1992; (86):181–184.

Blakemore SJ. Somatosensory activations during the observation of touch and a case of vision-touch synaesthesia. Brain. 2005; 128(7):1571–1583. doi:10.1093/brain/awh500. [PubMed: 15817510]

Botvinick M, Jha AP, Bylsma LM, Fabian SA, Solomon PE, Prkachin KM. Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. NeuroImage. 2005; 25(1):312–319. doi:10.1016/j.neuroimage.2004.11.043. [PubMed: 15734365]

Brambring M, Asbrock D. Validity of false belief tasks in blind children. Journal of Autism and Developmental Disorders. 2010; 40(12):1471–1484. doi:10.1007/s10803-010-1002-2. [PubMed: 20379770]

Bull Kovera M, Park RC, Penrod SD. Jurors' Perceptions of Eyewitness and Hearsay Evidence. Minn L Rev. 1992; 76:703.

Butts DA, Weng C, Jin J, Yeh C-I, Lesica NA, Alonso J-M, Stanley GB. Temporal precision in the neural code and the timescales of natural vision. Nature. 2007; 449(7158):92–95. doi:10.1038/nature06105. [PubMed: 17805296]

Calvo-Merino B, Glaser DE, Grèzes J, Passingham RE, Haggard P. Action observation and acquired motor skills: an FMRI study with expert dancers. Cerebral Cortex. 2005; 15(8):1243–1249. doi:10.1093/cercor/bhi007. [PubMed: 15616133]

Calvo-Merino B, Grèzes J, Glaser DE, Passingham RE, Haggard P. Seeing or Doing? Influence of Visual and Motor Familiarity in Action Observation. Current Biology. 2006; 16(19):1905–1910. doi:10.1016/j.cub.2006.07.065. [PubMed: 17027486]

Cox DD, Savoy RL. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage. 2003; 19(2):261–270. doi:10.1016/S1053-8119(03)00049-1. [PubMed: 12814577]

de Villiers PA. JW A, JA B. The Role of Language in Theory-of-Mind Development: What Deaf Children Tell Us. Why language matters for theory of mind. 2005

de Villiers JG, Garfield J, Gernet-Girard H, Roeper T, Speas M. Evidentials in Tibetan: Acquisition, semantics, and cognitive development. New Directions for Child and Adolescent Development. 2009; 2009(125):29–47. doi:10.1002/cd.248. [PubMed: 19787778]

Dehaene S, Cohen L. Cultural Recycling of Cortical Maps. Neuron. 2007; 56(2):384–398. doi:10.1016/j.neuron.2007.10.004. [PubMed: 17964253]

DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends in Cognitive Sciences. 2007; 11(8):333–341. doi:10.1016/j.tics.2007.06.010. [PubMed: 17631409]

Faller, M. Doctoral Dissertation. Stanford University; 2002. Semantics and Pragmatics of Evidentials in Cuzco Quechua..

Fedorenko E, Nieto-Castañón A, Kanwisher N. Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. Neuropsychologia. 2012; 50(4):499–513. doi:10.1016/j.neuropsychologia.2011.09.014. [PubMed: 21945850]

Figueras-Costa B, Harris P. Theory of mind development in deaf children: A nonverbal test of false-belief understanding. Journal of Deaf Studies and Deaf Education. 2001; 6(2):92. [PubMed: 15451854]

Fitneva SA. Epistemic marking and reliability judgments: Evidence from Bulgarian. Journal of Pragmatics. 2001; 33:401–420.

Formisano E, Kim DS, Di Salle F, van de Moortele PF, Ugurbil K, Goebel R. Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron. 2003; 40(4):859–869. [PubMed: 14622588]

Gallese V, Goldman A. Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences. 1998; 2(12):493–501. [PubMed: 21227300]

Gallese V, Sinigaglia C. What is so special about embodied simulation? Trends in Cognitive Sciences. 2011; 15(11):512–519. doi:10.1016/j.tics.2011.09.003. [PubMed: 21983148]

Gazzola V, Keysers C. The Observation and Execution of Actions Share Motor and Somatosensory Voxels in all Tested Subjects: Single-Subject Analyses of Unsmoothed fMRI Data. Cerebral Cortex. 2009; 19(6):1239–1255. doi:10.1093/cercor/bhn181. [PubMed: 19020203]

Goldman AI. Interpretation Psychologized. Mind & Language. 1989; 4(3):161–185.

Goldman, AI. Simulating Minds:The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford University Press; USA: 2006.

Gopnik, AM.; Meltzoff, AN. Words, Thoughts, and Theories. MIT Press; Cambridge, MA: 1997.

Gopnik AM, Wellman HM. Why the Child's Theory of Mind Really Is a Theory. Mind & Language. 1992; 7(1-2):145–171. doi:10.1111/j.1468-0017.1992.tb00202.x.

Gordon RM. Folk Psychology as Simulation. Mind & Language. 1989; 1(2):158– 171. doi:10.1111/j.1468-0017.1986.tb00324.x.

Goris RLT, Op de Beeck HP. Neural representations that support invariant object recognition. Frontiers in Computational Neuroscience. 2009; 3:3. doi:10.3389/neuro.10.003.2009. [PubMed: 19242556]

Harris P. Checking our sources: the origins of trust in testimony. Studies in History and Philosphy of Science. 2002a; (33):315–333.

Harris, P. What do children learn from testimony?. In: Carruthers, P.; Stich, S.; Siegal, M., editors. The Cognitive Basis of Science. Cambridge University Press; 2002b.

Harris PL. From Simulation to Folk Psychology: The Case for Development. Mind & Language. 1992; 7(1-2):120–144. doi:10.1111/j.1468-0017.1992.tb00201.x.

Harris PL, Koenig MA. Trust in testimony: how children learn about science and religion. Child Development. 2006; 77(3):505–524. doi:10.1111/j.1467-8624.2006.00886.x. [PubMed: 16686784]

Haxby JV. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science. 2001; 293(5539):2425–2430. doi:10.1126/science.1063736. [PubMed: 11577229]

Haynes J-D, Rees G. Decoding mental states from brain activity in humans. Nature Reviews Neuroscience. 2006; 7(7):523–534. doi:10.1038/nrn1931.

Immordino-Yang MH, McColl A, Damasio H, Damasio A. Neural correlates of admiration and compassion. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(19):8021–8026. doi:10.1073/pnas.0810363106. [PubMed: 19414310]

Johanson, L.; Utas, B. Evidentials. Walter de Gruyter; 2000.

Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain. Nature Neuroscience. 2005; 8(5):679–685. doi:10.1038/nn1444.

Kamitani Y, Tong F. Decoding Seen and Attended Motion Directions from Activity in the Human Visual Cortex. Current Biology. 2006; 16(11):1096–1102. doi:10.1016/j.cub.2006.04.003. [PubMed: 16753563]

Keysers C, Gazzola V. Integrating simulation and theory of mind: from self to social cognition. Trends in Cognitive Sciences. 2007; 11(5):194–196. doi:10.1016/j.tics.2007.02.002. [PubMed: 17344090]

Kilner JM, Frith CD. Action observation: inferring intentions without mirror neurons. Current Biology. 2007; 18(1):R32–3. doi:10.1016/j.cub.2007.11.008. [PubMed: 18177711]

Kilner JM, Paulignan Y, Blakemore SJ. An Interference Effect of Observed Biological Movement on Action. Current Biology. 2003; 13(6):522–525. doi:10.1016/S0960-9822(03)00165-9. [PubMed: 12646137]

Koenig MAM, Clément FF, Harris PLP. Trust in testimony: children's use of true and false statements. Psychological Science. 2004; 15(10):694–698. doi:10.1111/j.0956-7976.2004.00742.x. [PubMed: 15447641]

Koster-Hale J, Saxe R, Dungan J, Young LL. Decoding moral judgments from neural representations of intentions. Pnas. 2013:1–14. doi:10.1073/pnas.1207992110/-/DCSupplemental.

Kriegeskorte N, Bandettini P. Analyzing for information, not activation, to exploit high-resolution fMRI. NeuroImage. 2007; 38(4):649–662. doi:10.1016/j.neuroimage.2007.02.022. [PubMed: 17804260]

Landau, B.; Gleitman, LR. Language and Experience. Harvard University Press; 1985.

Lenci A, Baroni M, Cazzolli G, Marotta G. BLIND: a set of semantic feature norms from the congenitally blind. Behavior Research Methods. 2013:1–16. doi:10.3758/s13428-013-0323-4. [PubMed: 23055156]

Lombardo MV, Chakrabarti B, Bullmore ET, Baron-Cohen S, Consortium MA. Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. NeuroImage. 2011; 56(3):1832–1838. doi:10.1016/j.neuroimage.2011.02.067. [PubMed: 21356316]

Mahon BZ, Caramazza A. Judging semantic similarity: an event-related fMRI study with auditory word stimuli. Nsc. 2010; 169(1):279–286. doi:10.1016/j.neuroscience.2010.04.029.

Marmor GS. Age at onset of blindness and the development of the semantics of color names. Journal of Experimental Child Psychology. 25(2):267–278. n.d. [PubMed: 660091]

Mason RA, Just MA. Differentiable cortical networks for inferences concerning people's intentions versus physical causality. Human Brain Mapping. 2011; 32(2):313–329. doi:10.1002/hbm.21021. [PubMed: 21229617]

McAlpine L, Moore CD. The development of social understanding in children with visual impairments. Journal of Visual Impairment and Blindness. 1995; (89):349–358.

Meltzoff, AN. The case for a developmental cognitive science: Theories of people and things.. In: Brenmner, G.; Slater, A., editors. Theories of infant development. Blackwell; Oxford: 2004. p. 145-173.

Meltzoff AN. "Like me": a foundation for social cognition. Developmental Science. 2007; 10(1):126–134. doi:10.1111/j.1467-7687.2007.00574.x. [PubMed: 17181710]

Meltzoff AN, Brooks R. Flom R, Lee K, Muir D. Eyes wide shut: The importance of eyes in infant gaze following and understanding of other minds. Gaze Following: its development and signficance. 2007:217–241. [Eyes wide shut: The importance of eyes in infant gaze following and understanding of other minds.: Eyes wide shut: The importance of eyes in infant gaze following and understanding of other minds.]

Meristo M, Falkman KW, Hjelmquist E, Tedoldi M, Surian L, Siegal M. Language access and theory of mind reasoning: Evidence from deaf children in bilingual and oralist environments. Developmental Psychology. 2007; 43(5):1156–1169. doi:10.1037/0012-1649.43.5.1156. [PubMed: 17723042]

Miene, P.; Bordiga, E.; Park, R. The Evaluation of Hearsay Evidence: A Social Psychological Approach.. In: Castellan, NJ., editor. Individual and Group Decision Making. Psychology Press; 1993.

Miene P, Park RC, Bordiga E. Juror Decision Making and the Evaluation of Hearsay Evidence. 76 Minn . L. Rev. 1992; (683)

Millar S. Spatial representation by blind and sighted children. Journal of Experimental Child Psychology. 1976; 21(3):460–479. [PubMed: 939949]

Minter M, Hobson RP, B M. Congenital visual impairment and "theory of mind.". British Journal of Developmental Psychology. 1998; (16):183–196.

Moeller M, Schick B. Relations between maternal input and theory of mind understanding in deaf children. Child Development. 2006; 77(3):751–766. [PubMed: 16686799]

Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. NeuroImage. 2011; 56(2):400–410. doi:10.1016/j.neuroimage.2010.07.073. [PubMed: 20691790]

Nichols S, Stich S, Leslie A, Klein D. Varieties of Off-Line Simulation. Theories of Theories of Mind. 1996

Norman K, Polyn S, Detre G, Haxby J. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences. 2006; 10(9):424–430. [PubMed: 16899397]

O'neill DK, Gopnik A. Young children's ability to identify the sources of their beliefs. Developmental Psychology. 1991; 27(3):390–397.

Papafragou, A.; Li, P. Evidential morphology and theory of mind.. Presented at the Proceedings from the 26th Annual Boston University Conference on Language Development; Somerville, MA. Cascadilla Press; 2001.

Park RC. A Subject Matter Approach to Hearsay Reform. 86 Mich. L. Rev. 1987; (51)

Peelen MV, Atkinson AP, Vuilleumier P. Supramodal Representations of Perceived Emotions in the Human Brain. Journal of Neuroscience. 2010; 30(30):10127–10134. doi:10.1523/JNEUROSCI. 2161-10.2010. [PubMed: 20668196]

Peelen MV, Wiggett AJ, Downing PE. Patterns of fMRI Activity Dissociate Overlapping Functional Brain Areas that Respond to Biological Motion. Neuron. 2006; 49(6):815–822. doi:10.1016/j.neuron.2006.02.004. [PubMed: 16543130]

Perner, J. Understanding the Representational Mind. MIT Press; 1993.

Peterson C, Wellman HM. From fancy to reason: scaling deaf and hearing children's understanding of theory of mind and pretence. British Journal of Developmental Psychology. 2010; 27(Pt 2):297–310. doi:10.1348/026151008X299728. [PubMed: 19998533]

Peterson, Siegal M. Representing Inner Worlds: Theory of Mind in Autistic, Deaf, and Normal Hearing Children. Psychological Science. 1999; 10(2):126–129. doi:10.1111/1467-9280.00119.

Peterson JL, Webb J. Factors influencing the development of a theory of mind in blind children. British Journal of Developmental Psychology. 2000; 18(3):431–447. doi:10.1348/026151000165788.

Pyers J, Senghas A. Language Promotes False-Belief Understanding. Psychological Science. 2009; 20(7):805. [PubMed: 19515119]

Rabin JS, Gilboa A, Stuss DT, Mar RA, Rosenbaum RS. Common and unique neural correlates of autobiographical memory and theory of mind. Journal of Cognitive Neuroscience. 2010; 22(6): 1095–1111. doi:10.1162/jocn.2009.21344. [PubMed: 19803685]

Robinson EJ, Whitcombe EL. Children's Suggestibility in Relation to their Understanding about Sources of Knowledge. Child Development. 2003; 74(1):48–62. doi: 10.1111/1467-8624.t01-1-00520. [PubMed: 12625435]

Robinson EJ, Haigh SN, Nurmsoo E. Children's working understanding of knowledge sources: Confidence in knowledge gained from testimony. Cognitive Development. 2008; 23(1):105–118. doi:10.1016/j.cogdev.2007.05.001.

Rolls ET, Treves A. The neuronal encoding of information in the brain. Progress in Neurobiology. 2011; 95(3):448–490. doi:10.1016/j.pneurobio.2011.08.002. [PubMed: 21907758]

Rosel J, Caballer A, Jara P, Oliver JC. Verbalism in the Narrative Language of Children Who Are Blind and Sighted. Journal of Visual Impairment and Blindness. 2005; 99:413–425.

Samson D, Apperly IA, Braithwaite JJ, Andrews BJ, Bodley Scott SE. Seeing it their way: Evidence for rapid and involuntary computation of what other people see. Journal of Experimental Psychology: Human Perception and Performance. 2010; 36(5):1255–1266. doi:10.1037/a0018729. [PubMed: 20731512]

Saxe R. Against simulation: the argument from error. Trends in Cognitive Sciences. 2005; 9(4):174–179. doi:10.1016/j.tics.2005.01.012. [PubMed: 15808499]

Saxe R, Kanwisher N. People thinking about thinking people:: The role of the temporo-parietal junction in. NeuroImage. 2003; 19(4):1835–1842. [PubMed: 12948738]

Saxe R, Powell LJ. It's the thought that counts: specific brain regions for one component of theory of mind. Psychological Science. 2006; 17(8):692–699. doi:10.1111/j.1467-9280.2006.01768.x. [PubMed: 16913952]

Schick B, de Villiers P, de Villiers J, Hoffmeister R. Language and theory of mind: A study of deaf children. Child Development. 2007; 78(2):376–396. [PubMed: 17381779]

Serences JT, Boynton GM. Feature-Based Attentional Modulations in the Absence of Direct Visual Stimulation. Neuron. 2007; 55(2):301–312. doi:10.1016/j.neuron.2007.06.015. [PubMed: 17640530]

Shepard RN, Cooper LA. Representation of Colors in the Blind, Color-Blind, and Normally Sighted. Psychological Science. 1992; 3(2):97–104. doi:10.1111/j.1467-9280.1992.tb00006.x.

Singer T, Seymour Ben, O'Doherty J, Kaube H, Dolan RJ, Frith CD. Empathy for pain involves the affective but not sensory components of pain. Science. 2004; 303(5661):1157–1162. doi:10.1126/science.1093535. [PubMed: 14976305]

Smirnova A. Evidentiality in Bulgarian: Temporality, Epistemic Modality, and Information Source. Journal of Semantics. 2012 doi:10.1093/jos/ffs017.

Smith AT, Kosillo P, Williams AL. The confounding effect of response amplitude on MVPA performance measures. NeuroImage. 2011; 56(2):525–530. doi:10.1016/j.neuroimage.2010.05.079. [PubMed: 20566321]

Spunt R, Satpute A, Lieberman M. Identifying the what, why, and how of an observed action: an fMRI study of mentalizing and mechanizing during action observation. Journal of Cognitive Neuroscience. 2010; 23(1):63–74. [PubMed: 20146607]

Stich S, Nichols S. Folk Psychology: Simulation or Tacit Theory? Mind & Language. 1992; 7(1-2): 35–71.

Urwin, C. Dialogue and cognitive functioning in the early language development of three blind children.. In: Mills, AE., editor. Language acquisition in the blind child: Normal and deficient. College-Hill Press; San Diego, CA: 1983. p. 142-161.

Woolfe T, Want S. Signposts to development: Theory of mind in deaf children. Child Development. 2002; 73(3):768–778. [PubMed: 12038550]

- • We use multivoxel pattern analysis to explore the neural basis of theory of mind.

- • We decode the perceptual source (hearing vs. seeing) of others' mental states.

- • These representations of perceptual source are preserved in congenitally blind adults.

- • Representations of others' mental states are not fundamentally derived from first-person experiences.
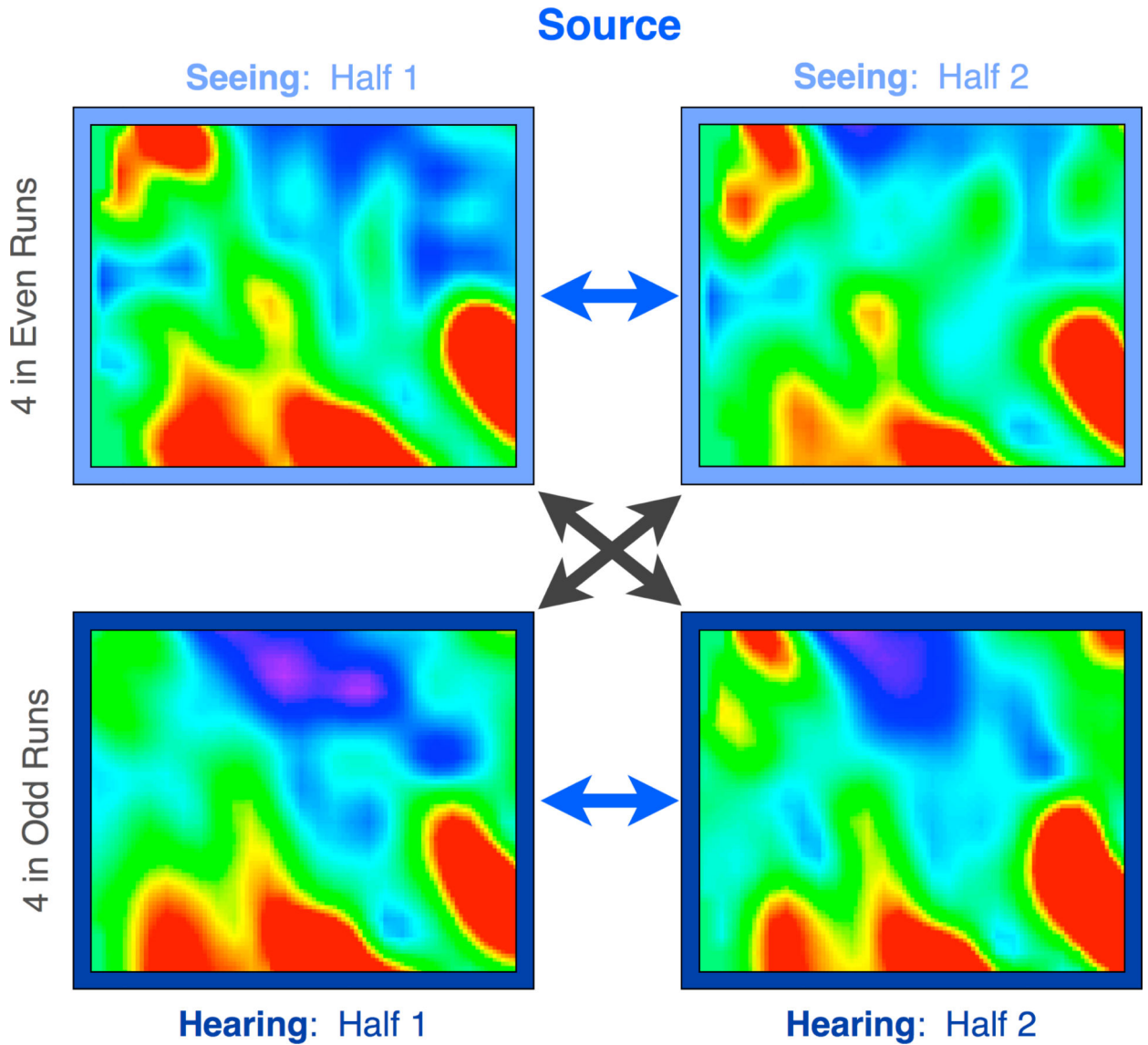
Figure 1.
Sample stimuli.

Figure 2. Split half MVPA analysis
The data from each condition (e.g. hearing and seeing) were divided by run, and for each participant, we asked whether the within condition correlations (blue arrows) were higher than the across conditions correlations (grey arrows). Data from one blind participant: within-condition correlation = 1.3; across-condition correlation = 0.8.
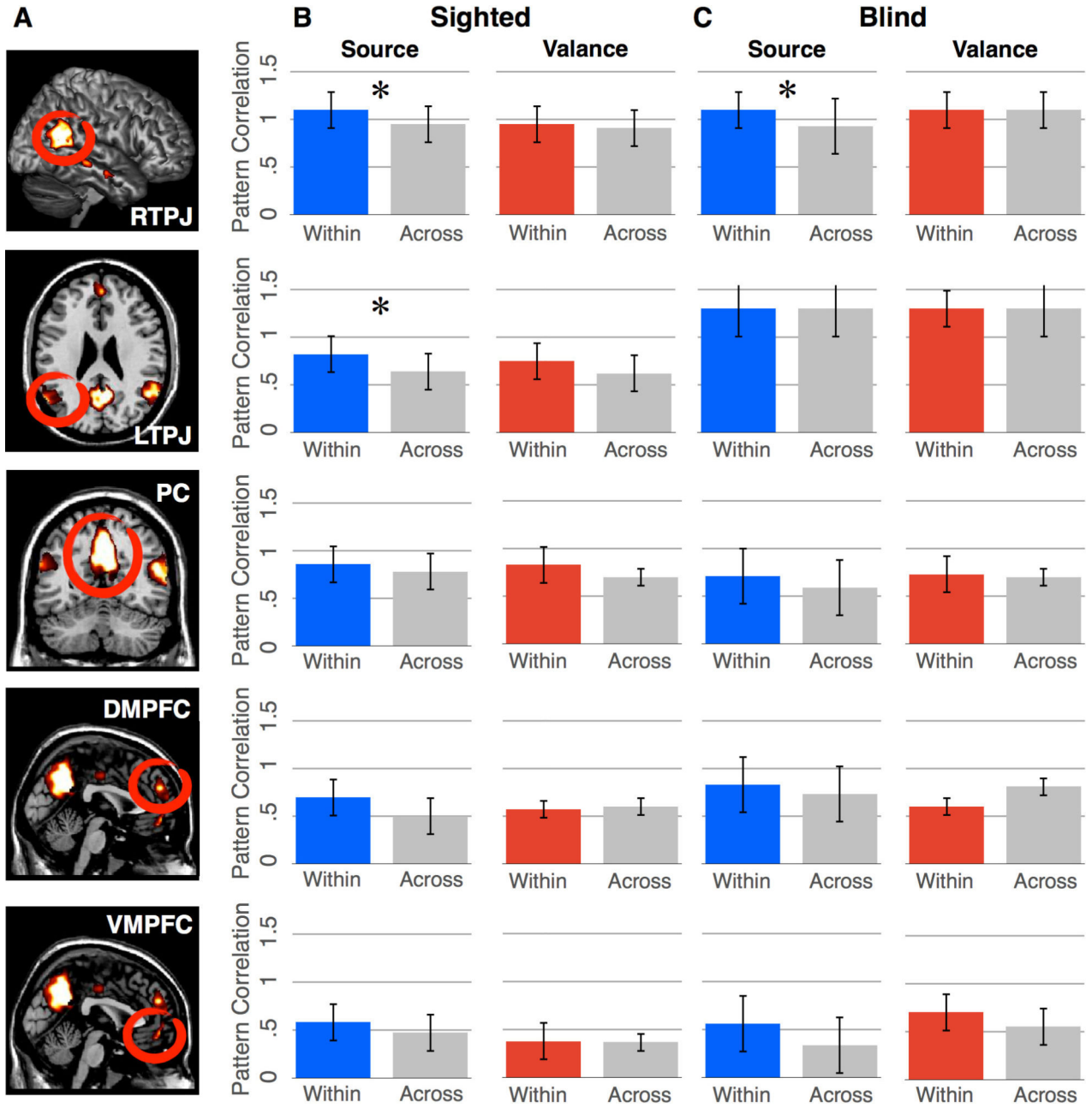
Figure 3. MVPA results

(A) Canonical theory of mind brain regions; for all analyses, theory of mind brain regions were individual defined in each participant. **(B)** Sighted adults (n=13) show pattern discrimination for beliefs based on seeing vs. hearing in the right and left TPJ, but not in any other theory of mind region. **(C)** These representations of seeing and hearing persist in the RTPJs of blind adults (n=9).

Table 1

Z scores for *within* versus *across* condition comparisons of *seeing* and *hearing*.

	Region	Within	Across	Significance
Sighted	**RTPJ**	1.1±0.2	0.95±0.2	**t(12)=2.0, p=0.03***
	LTPJ	0.82±0.2	0.64±0.2	**t(11)=2.0, p=0.04***
	PC	0.86±0.2	0.78±0.2	t(12)=0.93, p=0.19
	DMPFC	0.7±0.1	0.5±0.2	t(10)=1.2, p=0.12
	VMPFC	0.58±0.2	0.47±0.2	t(10)=0.75, p=0.23
Blind	**RTPJ**	1.1± 0.2	0.93±0.3,	**t(7)=2.0, p=0.04***
	LTPJ	1.3±0.3	1.3±.03	t(8)=0.2 p=0.4
	PC	0.71±0.2	0.59±0.2	t(8)=1.3, p=0.12
	DMPFC	0.83±0.2	0.73±0.1	t(8)=1.4, p=0.11
	VMPFC	0.56±0.2	0.34±0.2	t(8)=1.4, p=0.11