Right Temporoparietal Junction Encodes Inferred Visual Knowledge of Others

Authors: Branden J. Bio*[1], Arvid Guterstam*[1,2,3], Mark Pinsk[4], Andrew I. Wilterson[1], Michael S. A. Graziano[1,4]

*Equal contribution

1. Department of Psychology, Princeton University, Princeton, NJ 08544

2. Department of Clinical Neuroscience, Karolinska Institutet, 171 77 Solna, Stockholm, Sweden

3. Department of Neurology, Karolinska University Hospital, Stockholm, Sweden.

4. Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544

Corresponding author: Michael Graziano**,** Email: graziano@princeton.edu

Abstract

When people make inferences about other people's minds, called theory of mind (ToM), a cortical network becomes active. The right temporoparietal junction (TPJ) is one of the most consistently responsive nodes in that network. Here we used a pictorial, reaction-time, ToM task to study brain activity in the TPJ and other cortical areas. Subjects were asked to take the perspective of a cartoon character and judge its knowledge of a visual display in front of it. The right TPJ showed evidence of encoding information about the implied visual knowledge of the cartoon head. When the subject was led to believe that the head could see a visual change take place, activity in the right TPJ significantly reflected that change. When the head could apparently not see the same visual change take place, activity in the right TPJ no longer significantly reflected that change. The subject could see the change in all cases; the critical factor that affected TPJ activity was whether the subject was led to think the cartoon character could see the change. We also found that whether the beliefs attributed to the cartoon head were true or false did not significantly affect activity in the present paradigm. These results suggest that the right TPJ may play a role in modeling the contents of the minds of others, perhaps more than it participates in evaluating the truth or falsity of that content.

Keywords: False belief; fMRI; Mentalizing; Temporoparietal Junction; Theory of Mind

Introduction

Building a model of other people's thoughts, emotions, and beliefs, also called theory of mind (ToM), is foundational to our social lives (Wimmer and Perner 1983; Baron-Cohen 1997; Frith and Frith 2003; Wellman 2018). A large literature shows that ToM tasks tend to activate a specific network of areas in the human cerebral cortex (Fletcher et al. 1995; Gallagher et al. 2000; Vogeley et al. 2001; Gallagher and Frith 2003; Saxe and Kanwisher 2003; Frith and Frith 2006; Saxe 2006; Gobbini et al. 2007; Spreng et al. 2009; Mar 2011; Kelly et al. 2014; Schurz et al. 2014; van Veluw and Chance, 2014; Igelström et al. 2016; Richardson et al. 2018). One of the most consistently activated areas is the temporoparietal junction (TPJ), sometimes bilaterally but with a bias toward the right side. Other areas include the superior temporal sulcus (STS), again sometimes bilaterally but with a bias toward the right side; the medial prefrontal cortex (MPFC); and the precuneus. Other brain areas are also reported in ToM studies, such as the temporal pole and the amygdala, but the areas listed above are often more consistently active during ToM tasks, as shown in meta-analysis studies (Mar 2011; Schurz et al. 2014; van Veluw and Chance 2014).

Experiments on the ToM cortical network often use a classic paradigm called the false belief task (Wimmer and Perner 1983; Baron-Cohen et al. 1985). In it, the subject of the experiment must decide, based on information known to be available to a character in a story, whether the character thinks that A or B is true. For example, if Sally originally put her sandwich into box A, will she still think it is in box A, even after someone else, unbeknownst to Sally, has moved it to box B? To answer correctly, the subject of the experiment must have a sophisticated

enough theory of Sally's mind to realize that Sally can believe something that is false – that is distinct from the world around her.

The purpose of the present study was to use a modified version of the false belief task to compare two specific hypotheses. In a recent behavioral study, we designed a pictorial variant of the false belief task (Bio et al. 2018). Visual ToM tasks incorporating cartoons or videos have been used before (Gallagher et al. 2000; Grèzes et al. 2004; Marjoram et al. 2006; Hooker et al. 2008; Sommer et al. 2010; Rothmayr et al. 2011; Richardson et al. 2018). Our version was designed to build up a set of information over the course of each trial, leading to a final decision that participants must make. As shown in Figure 1A, first, two cartoon heads appeared, looking at two open boxes. Second, the "ball," a red dot, appeared in one box. Third, one head was shown having its eyes covered such that it could no longer see what was in either box, while the other head remained with uncovered eyes. Fourth, on half of the trials, the ball switched from one box to the other. Fifth and finally, a question mark appeared in one head or the other. The participants were required to decide whether the head indicated by the question mark "believed" the ball to be in box 1 or box 2. The purpose of this incremental presentation of information, and the use of two heads, one covered and the other uncovered, was to ensure that the subjects would not know how to answer the question until the last pictorial piece of information, the question mark indicating the correct head, was presented. At that moment, subjects had all necessary information to make the judgment, and they responded in a speeded manner within a limited response window (1.5 s). The design therefore converted the false belief task into a pictorial, reaction-time task, in which a single event (the appearance of the question mark) triggered the moment when subjects needed to make a ToM judgment. With all stimuli being fully right-

versus-left counterbalanced across trials, the conditions were nearly visually identical and provided a good control for each other.

We conceptualized the paradigm as a 2X2 design, as shown in Figure 1B. The first variable was whether the indicated cartoon head had its vision blocked. In half the trials, the subjects had to judge the visual beliefs of the cartoon head whose eyes were covered, whereas in the other half of trials, the subjects had to judge the visual beliefs of the cartoon head whose eyes were uncovered. The second variable was whether the ball switched from one box to the other. In half the trials, the ball switched boxes midway through the trial, whereas in the other half of trials, the ball began in one box and remained there without switching. This design resulted in four trial conditions that we termed blocked-switched (BS), blocked-nonswitched (BnS), nonblocked-switched (nBS), and nonblocked-nonswitched (nBnS).

In the present experiment, subjects performed this pictorial ToM task in a magnetic resonance imaging (MRI) scanner, which measured brain activity evoked by the subjects' decisions on each trial. The experiment was designed to test two hypotheses. The hypotheses are formulated with respect to the right TPJ, because that cortical area is most consistently active during ToM and false belief tasks. However, the same hypotheses could also apply to other nodes in the ToM network.

Hypothesis 1

We hypothesized that, when asked to judge the cartoon character's belief about ball location, the subjects will reconstruct the cartoon's general visual knowledge of the scene, and the right TPJ will show evidence of encoding that visual knowledge. In this hypothesis, activity in the right TPJ should distinguish between two conditions in particular: nBS and nBnS. In these

two conditions, the cartoon can see the ball at all times. In one trial type (nBnS), it can see the ball placed in one box, and see that it stays there. In the other trial type (nBS), the head can see the ball placed in one box, and can see that it switches to the other box. The cartoon head therefore has two, different knowledge sets about the visual display in front of it. In hypothesis 1, the subject, tasked with assessing the head's perspective on that stimulus display, reconstructs that the head has different knowledge in the two conditions. Brain areas that reconstruct the knowledge of the head should show a difference in activity. In this same hypothesis, however, activity in the right TPJ should not distinguish between the BS and the BnS conditions. In these two conditions, the cartoon can see the ball placed in a box at the start of the trial and then its eyes are covered. It cannot see whether the ball is switched or not. From the perspective of the head, the BS and the BnS trial types are the same. The cartoon head therefore has the same knowledge about the visual display in front of it, in both trial types. The subject, tasked with assessing the head's perspective on that stimulus display, should reconstruct that the head has the same knowledge in the two conditions. Brain areas that reconstruct the knowledge of the head should not show a difference in activity.

In specific, hypothesis 1 predicts two significant differences in relation to right TPJ activity. First, we should find a significant difference between nBnS and nBS trials. Second, the nBnS-versus-nBS contrast should be significantly greater than the BS-versus-BnS contrast. The BS-versus-BnS contrast serves as a control to ensure that the results are not simply caused by the subject seeing the ball switch boxes. In both the nBnS-versus-nBS contrast, and the BS-versus-BnS contrast, the subject can see the ball switch boxes. But only in the nBnS-versus-nBS contrast does the subject realize that the cartoon head also sees the ball switch boxes. Thus,

activity that follows the predictions of hypothesis 1 would reflect the subject's reconstruction of the cartoon head's visual knowledge.

Hypothesis 1 depends on the subjects using ToM reasoning to solve all four conditions in the task. In at least one traditional view, only false belief trials, not true belief trials, require ToM reasoning. For example, it could be argued that when the cartoon face is unblocked (and thus can see the ball), the subject does not need to use ToM reasoning to solve the task, and can simply state which box the ball is actually in, ignoring the perspective of the face altogether. However, a strategy of ignoring the face would work only for true belief trials and leave the subjects with poor scores on false belief trials. To ignore the face on true belief trials, and yet also consider the perspective of the face on false belief trials, would require distinguishing the true from the false belief trials, which would require understanding whether the cartoon character has a true or false belief, which would require using ToM reasoning. We suggest, therefore, that because subjects scored with high accuracy on all trial types, they must have used some degree of ToM reasoning in all trial types. Moreover, it has been argued that even when ToM reasoning is not logically required to solve a task, as long as the option for it exists, people automatically use it (Saxe and Kanwisher 2003). In constructing our paradigm, therefore, we assumed that all trial types, whether true or false belief, recruited ToM reasoning.

Hypothesis 1 also depends on the ability of the experiment to measure extremely subtle differences that are likely to be small in magnitude. Past ToM experiments using brain scanning often used a block design, relying on an average of brain activity across many seconds as subjects performed a continuous task such as reading a story (Fletcher et al. 1995; Gallagher et al. 2000; Vogeley et al. 2001; Saxe and Kanwisher 2003; Gobbini et al. 2007; Lee at al. 2011). Moreover, past experiments also often studied the difference between social cognition trials and

entirely non-social trials, providing a large difference in cognitive conditions. In the present experiment, we analyzed brain activity evoked at the time of the ToM decision, within the brief, 1.5 s reaction-time window. We also tested subtle differences between nearly identical trials, all of which probably engaged ToM reasoning. This approach has both a benefit and a cost. What we gain in the ability to target specific hypotheses about ToM reasoning, we lose in the likely magnitude of the effect. The result therefore depends on the sensitivity of the measurement and the analysis technique, discussed further below.

Hypothesis 2

Hypothesis 2 predicts that activity in the right TPJ will distinguish between false belief and true belief trials.

Many previous studies have compared false belief to true belief conditions, on the suggestion that false belief reasoning might require especially complex or intensive ToM, or might be processed in a specific part of the ToM network as distinct from true belief reasoning (Hooker et al. 2008; Aichhorn et al. 2009; Sommer et al. 2010; Döhnel et al. 2012). Other have suggested that ToM reasoning is used robustly whether a trial type includes false or true beliefs (Saxe and Kanwisher 2003). The results of these many previous studies are mixed. Though there is evidence of false belief processing emphasized in some subregions of the right TPJ (Aichhorn et al. 2009; Sommer et al. 2010; Döhnel et al. 2012), other researchers have argued that false and true belief conditions do not result in measurably different activity in the right TPJ (Saxe and Kanwisher 2003). Related to this hypothesis, it has been suggested (Mitchell 2009) that the right TPJ may be involved in filtering out or ignoring what is actually happening right now in one's

own experience, and instead building a model of what could, hypothetically, be happening in another mind.

In hypothesis 2, the activity in the right TPJ should distinguish between the BS and the BnS conditions. In these two conditions, the cartoon can see the ball placed in a box at the start of the trial and then its eyes are covered. It cannot see whether the ball is switched or not. When the ball is not switched (BnS) the head should have a true belief about the ball's location, and when the ball is switched (BS) the head should have a false belief. If the TPJ encodes the truth status of the cartoon's beliefs, then its activity should distinguish between those two conditions. In this same hypothesis, however, activity in the right TPJ should not distinguish between the nBS and the nBnS conditions. In these two conditions, the cartoon can see the ball at all times, whether it switches or not. The cartoon has only true beliefs, no false beliefs. If the TPJ encodes the truth versus falsity of the cartoon's beliefs, then its activity should not distinguish between those two conditions.

In specific, hypothesis 2 predicts two significant differences in relation to right TPJ activity. First, we should find a significant difference between BnS and BS trials. Second, the BnS-versus-BS contrast should be significantly greater than the nBS-versus-nBnS contrast. Hypothesis 2 therefore predicts exactly the opposite pattern of results as hypothesis 1.

Materials and Methods

Subjects

All subjects provided informed consent and all procedures were approved by the Princeton Institutional Review Board. We tested 28 healthy human volunteers (17 females, 27

right-handed, aged 18-50, normal or corrected to normal vision). Subjects were recruited from a paid subject pool, receiving 40 USD for participation.

Experimental setup

Before scanning, all participants received task instructions and completed practice trials on a laptop computer outside of the MRI scanner. During scanning, subjects laid in a supine position on the MRI bed and used an angled mirror mounted on top of the head coil to view a screen approximately 80 cm from the eyes, on which visual stimuli were projected using a digital light processing projector (Hyperion MRI Digital Projection System, Psychology Software Tools, Sharpsburg, PA, USA) with a resolution of 1920 x 1080 pixels at 60 Hz. A PC running MATLAB (MathWorks, Natick, MA, USA) and the Psychophysics Toolbox (Brainard 1997) were used to present visual stimuli. A 5-button response unit (Psychology Software Tools Celeritas, Sharpsburg, PA, USA) was strapped to the subjects' dominant hand. Subjects used only the index and middle fingers to indicate responses.

Behavioral task

The task events are illustrated in Figure 1A. Participants saw a cartoon that included two heads, two boxes, and a ball. The ball was located in one of two boxes and the participant had to decide whether a cartoon head would most likely believe the ball to be in box 1, to the left, or box 2, to the right. Participants responded by button press only at the end of the trial when one of the two cartoon heads was indicated as the target for the ToM judgment.

Each trial began with a black fixation cross at the center of a white background. Participants were instructed to fixate on the cross. After 500 ms, the fixation cross was joined by

a top-down view of two cartoon heads and two numbered boxes. The heads were centered 3.25 degrees to the left (head 1) and right (head 2) of the vertical midline of the screen, positioned on the horizontal midline (at the same height as the fixation cross). The boxes were centered 12 degrees to the left (box 1) and right (box 2) of the midline, and 9 degrees above the horizontal midline. After another 500 ms, a red ball appeared in one of the two boxes (half of the trials in box 1, half of the trials in box 2). Participants had been told in the instruction period that, in this configuration, both heads could see where the ball was located. After 1000 ms, one of the heads was blocked with a curved partition directly in front of it (half of the trials blocking head 1, half of the trials blocking head 2). Participants had been told in the instruction period that the blocked head could no longer see either the boxes or the ball, but that the other head could still see everything as before.

In half of the trials, 1000 ms after the blocking partition appeared, the ball switched position to the opposite box. If it was initially in box 1, it moved to box 2; if it was initially in box 2, it moved to box 1. The head that was blocked should therefore "believe" the ball to be still in the original box, and the head that was unblocked should "see" the ball move to the new box. In the other half of trials, the ball did not switch positions.

Finally, 4000 ms after the start of the trial, a question mark appeared inside one of the heads (half of trials in head 1, half of trials in head 2). The question mark indicated which head was to be the target of the participant's judgment. The participant was instructed to respond as quickly as possible once the question mark appeared. By pressing one of two buttons on the button box, the participant reported whether the indicated head would most likely think the ball was in box 1 or box 2. Participants were allowed a response window of 1500 ms. Trials on which participants exceeded the given time to respond were not included in the analysis. Participants responded within the correct time window on most trials

(98%). After the response window, the display of heads and boxes disappeared and a variable, 1000 - 3000 ms inter-trial interval followed, after which the next trial began with the onset of the fixation cross.

In summary, the task included the following conditions: the red dot could be initially presented in box 1 or box 2; the blocking screen could be placed in front of the left or right head; the red dot could be switched to the opposite box or remain in the same box; and the question mark could be presented in the left or right head. This 2X2X2X2 design resulted in 16 trial types, presented in a counterbalanced and randomized order. The trial types were collapsed into four main conditions for purposes of analysis (see Figure 1B). These conditions formed a 2 X 2 design as follows: blocked trials, on which the head indicated by the question mark was blocked by the screen, versus nonblocked trials, on which the indicated head was not blocked by the screen; and switched trials, on which the ball moved to the opposite box, versus nonswitched trials, on which the ball remained in the initial box. As shown in Figure 1B, these four conditions were labeled as blocked switched (BS), blocked nonswitched (BnS), nonblocked switched (nBS), and nonblocked nonswitched (nBnS).

Participants performed 256 trials (64 per main condition), in 8 runs of 32 trials each. Each run took approximately 5.5 minutes to complete and included 5 s of baseline before the onset of the first trial and 10 s of baseline after the offset of the last trial.

fMRI data acquisition

Functional imaging data were collected using a 3T MAGNETOM Skyra (Siemens Healthineers AG, Erlangen, Germany) scanner equipped with a 64-channel head/neck coil. Gradient-echo T2*-weighted echo-planar images (EPI) with blood-oxygen dependent (BOLD) contrast were used as an index of brain activity (Logothetis et al. 2001). Functional image

volumes were composed of 46 near-axial slices with a thickness of 3.0 mm (with no interslice gap), which ensured that the entire brain excluding the cerebellum was within the field-of-view in all subjects (80 x 80 matrix, 2.5 mm x 2.5 mm in-plane resolution, TE = 30 ms, flip angle = 75°). Simultaneous multi-slice (SMS) imaging was used (SMS factor = 2). One complete volume was collected every 1.5 s (TR = 1500 ms). A total of 1300 functional volumes were collected for each participant, divided into 8 runs (130 volumes per run). The first five volumes of each run were discarded to account for non-steady-state magnetization. A high-resolution structural image was acquired for each participant at the end of the experiment (3D MPRAGE sequence, voxel size = 1 mm isotropic, FOV = 256 mm, 176 slices, TR = 2300 ms, TE = 2.96 ms, TI = 1000 ms, flip angle = 9°, iPAT GRAPPA = 2). At the end of each scanning session, matching spin echo EPI pairs were acquired with reversed phase-encode blips, resulting in pairs of images with distortions going in opposite directions for blip-up/blip-down susceptibility-derived distortion correction.

FMRI preprocessing

Results included in this manuscript come from preprocessing performed using FMRIPREP version 1.2.3 (Esteban et al. 2019) (RRID:SCR_016216), a Nipype (Gorgolewski et al. 2011) (RRID:SCR_002502) based tool. Each T1w (T1-weighted) volume was corrected for INU (intensity nonuniformity) using N4BiasFieldCorrection v2.1.0 (Tustison et al. 2010) and skull-stripped using antsBrainExtraction.sh v2.1.0 (using the OASIS template). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al. 2009) (RRID:SCR_008796) was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0 (Avants et al. 2008) (RRID:SCR_004757), using brain-

extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (Zhang et al. 2001) (FSL v5.0.9, RRID:SCR_002823).

Functional data was slice time corrected using 3dTshift from AFNI v16.2.07 (Cox 1996) (RRID:SCR_005927) and motion corrected using mcflirt (FSL v5.0.9) (Jenkinson et al. 2002). This procedure was followed by co-registration to the corresponding T1w using boundary-based registration (Greve and Fischl 2009) with six degrees of freedom, using flirt (FSL). Motion correcting transformations, BOLD-to-T1w transformation and T1w-to-template Montreal Neurological Institute (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation.

Many internal operations of FMRIPREP use Nilearn (Abraham et al. 2014) (RRID:SCR_001362]) principally within the BOLD-processing workflow. For more details of the pipeline see https://fmriprep.readthedocs.io/en/latest/workflows.html.

MVPA analysis

We analyzed the data using MVPA, which tests whether patterns of brain activity can be used to decode the distinction between two conditions. It is a more sensitive analysis than the more common, simple univariate subtraction methods, and the study was designed from the outset to use MVPA (thus many trials per condition were included). Two independent MVPA comparisons were performed: BS versus BnS trials, and nBS versus nBnS trials. We tested both comparisons within a set of six ROIs.

The ROIs were defined as spheres centered on the statistical peaks reported in an activation likelihood estimation meta-analysis of 16 fMRI studies (including 291 subjects)

involving ToM reasoning (van Veluw and Chance 2014), in accordance with the approach used in Guterstam et al. (2021) and the generally accepted guidelines in ROI analysis (Poldrack, 2007). The ROIs are shown in Figure 2. The peaks were located in six areas: the left TPJ (Montreal Neurological Institute [MNI]: -52, -56, 24), right TPJ (MNI: 55, -53, 24), left STS (MNI: -59, -26, -9), right STS (MNI: 59, -18, -17), MPFC (MNI: 1, 58, 19), and the precuneus (MNI: -3, -56, 37). The radius of the ROI spheres was 10 mm, corresponding to the approximate volume (4,000 mm^3) of the largest clusters (TPJ and MPFC) reported in the meta-analysis study used here to define the ROIs (van Veluw and Chance 2014). The same sphere radius was used for all ROIs.

The fMRI data from all participants were analyzed with the Statistical Parametric Mapping software (SPM12) (Wellcome Department of Cognitive Neurology, London, UK) (Friston et al. 1994). We first used a conventional general linear model (GLM) to estimate regression beta coefficients for each individual trial (i.e., 256 regressors), focusing on the phase of each trial over 1.5 s immediately after the question mark appeared (the time window in which the subjects were allowed to judge and make a response). A regressor of no interest modeled the initial 4 s of the trial across all conditions. Each regressor was modeled with a boxcar function and convolved with the standard SPM12 hemodynamic response function. In addition, 8 run-specific regressors controlling for baseline differences between runs, and six motion regressors, were included. The trial-wise beta coefficients (i.e., 256 beta maps) were then submitted to subsequent multivariate analyses (Haxby et al. 2001).

The MVPA was carried out using The Decoding Toolbox (TDT) version 3.999 (Hebart et al. 2015) for SPM. For each subject and ROI, we used linear support vector machines (SVMs, with the fixed regularization parameter of C = 1) to compute decoding accuracies. To ensure

independent training and testing data sets, we used a leave-one-run-out cross-validation approach. For each fold, an SVM was then trained to discriminate activity patterns belonging to the contrasted trial types in seven runs, and then tested on the trials in the left-out run, repeated for all runs, resulting in a run-average decoding accuracy for each ROI and subject.

For statistical inference, the true group mean decoding accuracy was compared to a null distribution of group mean accuracies obtained from permutation testing. The same MVPA was repeated within each subject and ROI using permuted condition labels (1000 iterations). A p value was computed as (1+the number of permuted group accuracy values > true value)/(1+the total number of permutations). To control for multiple comparisons across the six ROIs, we used the false discovery rate (FDR) correction (Benjamini and Hochberg 1995). In addition, we also computed a bootstrap distribution around the true group mean accuracy by resampling individual-subject mean accuracies with replacement (1000 iterations), from which a 95% confidence interval (CI) was derived (Nakagawa and Cuthill 2007).

Beyond the targeted hypotheses of this study concerning the six ROIs, we also used a whole-brain searchlight analysis (Kriegeskorte et al. 2006) to test for possible areas of decoding outside the ROIs. The searchlight analysis is conceptually different from the ROI analysis. It is not targeted to specific brain areas on the basis of predictions, and therefore is more statistically conservative because of brain-wide multiple comparisons correction. In general, one would not expect the searchlight analysis to align with the ROI analysis. It is possible to obtain significant results in the ROI analysis that do not appear in the searchlight analysis. Instead, the searchlight analysis is useful for revealing clusters of strong decoding in areas that were not anticipated by hypothesis.

For the searchlight analysis, first, the brain was partitioned into overlapping voxel clusters of spherical shape (10-mm radius). In each of these clusters, a decoding accuracy was computed using the same model input, SVM parameters, and procedures as described for the ROI analysis. For each contrast between two trial types, this process resulted in a decoding accuracy map for each subject, in which the value of each voxel represents the average proportion of correctly classified trials relative to chance level (50%) based on the 10 mm sphere of tissue surrounding that voxel. The subject-wise decoding maps were then smoothed using a 3-mm full-width-half-maximum Gaussian kernel, and entered into a second-level analysis using SPM12. In that analysis, for statistical inference, we employed a cluster-level, whole-brain approach to find clusters that passed the threshold of $p < 0.05$, corrected for brain-wide multiple comparisons using the family-wise error rate correction as implemented by SPM12.

Univariate analysis

We subjected the data to univariate analyses to control for potential univariate effects that could contribute to classifier performance in the MVPA. The preprocessed data was smoothed using a 6-mm full-width-half-maximum Gaussian kernel. In the first-level analysis, we modeled the data using the same approach as described above for the MVPA, but defined one regressor per experimental condition (as opposed to one regressor per trial). We then defined linear contrasts in the GLM, and the contrast images from all subjects were entered into a random effects group analysis. For statistical inference, we searched for clusters that passed the threshold of $p < 0.05$, corrected for multiple comparisons either within each of the six ROIs, or using the whole brain as search space, using the familywise error rate correction as implemented by SPM12.

Results

Task performance

Subjects performed the task at high levels of accuracy, suggesting that they understood the instructions and attributed beliefs to the cartoon heads as intended. (Overall accuracy, 94.8%; for BS trials, 91.6%; BnS trials, 95.2%; nBS trials, 96.0%; nBnS trials, 96.2%; overall lateny = 1006 ms; mean latency for BS trials = 1007 ms, SEM = 29; for BnS trials = 977 ms, SEM = 31; for nBS trials = 1035 ms, SEM = 22; for nBnS trials = 1005 ms, SEM = 28.)

ROI analysis

Figures 2 and Table 1 show the results for the ROI analysis, for the nBnS-versus-nBS contrast. In each panel, the red line shows the accuracy of the MVPA analysis in decoding which trial type occurred, compared to a chance level of 50%. The histogram shows the null distribution of decoding accuracies based on permutation testing with shuffled conditions labels (1000 iterations).

According to hypothesis 1, cortical areas in the ToM network, especially the right TPJ, should show significant decoding for the nBnS-versus-nBS contrast. The results do show a significant decoding for nBS versus nBnS trials in the right TPJ. The magnitude of decoding accuracy was 53%, compared to the chance level of 50%, but was highly statistically reliable, thus showing that some information about the nBS-versus-nBnS contrast was highly likely to be present in the right TPJ ($p = 0.004$; see also 95% confidence intervals in Table 1). Even when

corrected for multiple comparisons across the six defined ROIs, it remained statistically significant ($p = 0.024$ corrected using FDR). When subjects thought that the head could see the ball switch boxes, then the right TPJ was affected by the switch.

Figure 4 and Table 2 show the results for the ROI analysis, for the BnS-versus-BS contrast. None of the six ROIs showed any significant decoding (the decoding accuracy was not significantly different from the chance level of 50%; see Table 2 for p values and for 95% confidence intervals). We therefore did not find any evidence of a difference in the TPJ, or other ToM areas, between processing switched and nonswitched trial types when subjects thought that the head could not see the ball switch boxes.

Finally, we compared the strength of the decoding obtained in the nBnS-versus-nBS contrast and in the BnS-versus-BS contrast. The decoding in the right TPJ for nBS versus nBnS trials was significantly greater than the decoding in the right TPJ for BS versus BnS trials ($p = 0.0432$, permutation testing with 10,000 iterations).

Searchlight analysis

As a further exploration beyond the targeted hypotheses of this study, we used a whole-brain searchlight analysis (Kriegeskorte et al. 2006) to test for possible areas of decoding outside the ROIs. Because the searchlight analysis does not test strong *a priori* hypotheses and requires statistical correction across the full brain, it is much less sensitive. The searchlight comparison between nBS and nBnS trials revealed no significant areas of decoding at the brain-wide level; likewise, the searchlight comparison between BS and BnS trials revealed no significant areas of decoding at the brain-wide level.

Univariate analysis

To control for potential univariate effects that could drive classifier performance in the decoding analyses, we examined the bi-directional contrasts for the BS-versus-BnS and the nBS-versus-nBnS comparisons (i.e., BS > BnS, BnS > BS, nBS > nBnS, and nBnS > nBS). None of the contrasts revealed significant activity, neither within the ROIs nor at the whole-brain level. The same result was found for all six possible specific comparisons (12 contrasts) between individual conditions, the two main effects (4 contrasts), and the interaction (2 contrasts). The absence of any univariate effect within the ROIs, or anywhere else in the brain, confirm that the stimuli were well matched. These findings are compatible with previous studies (Hassabis et al. 2009) that demonstrated the superiority of pattern-sensitive multivariate analyses compared to conventional univariate approaches for detecting differences in activity between conditions with highly similar macroscopic characteristics.

Discussion

The present experiment used fMRI to measure brain activity during a pictorial, reaction-time, ToM task that incorporated both false belief and true belief trials. We used the task to test two specific hypotheses. In hypothesis 1, only when the head was unblocked, and by implication could see whether the ball switched or not, should the ToM network react differently to the switch and nonswitch conditions, reflecting a difference in visual knowledge attributed to the head. In hypothesis 2, brain areas in the ToM cortical network should respond differently to false belief trials and true belief trials. The two hypotheses predicted opposite activity patterns. The results supported hypothesis 1. Note that we cannot rule out hypothesis 2. The ToM brain areas may still encode the truth or falsity of other people's beliefs. Such a signal might be present but

too subtle to be measured by our paradigm. The results do, however, indicate that in our paradigm the right TPJ is significantly more sensitive to the implied contents of the cartoon's mind than it is to the truth or falsity of the cartoon's beliefs.

The subjects could see the ball switch from one box to another, and thus could see the difference between switch and nonswitch trials. Could the right TPJ have simply reacted to the difference between seeing a switched and a nonswitched trial? The data rule out this possibility. The activity difference between switch and nonswitch trials was seen only in trials when both the subjects and the cartoon head could see whether the switch took place (nBnS versus nBS), not on trials when the subject could see the switch and the cartoon head could not (BnW versus BS). If the right TPJ activity reflected a difference between switched and nonswitched trials, it was evidently not a general effect, but only occurred when the subjects thought that the cartoon character could see the switch take place. We suggest, therefore, that our interpretation in terms of modeling the mind states of others is the most plausible one.

One could argue that the results are extremely subtle (53% decoding accuracy in the right TPJ for the nBnS-versus-nBS comparison). However, the magnitude of the result is not at issue here. The effect was highly statistically significant, indicating that information about the nBnS-versus-nBS comparison was highly likely to be present in the right TPJ. One usefulness of the MVPA analysis method is its sensitivity to extremely subtle effects that reveal information present in brain activity. The design of the study (measuring differences between conditions that are extremely closely matched) could be seen as a strength of the study, allowing for targeted hypothesis testing.

The present results might help to explain the somewhat mixed results of previous studies that compared false belief and true belief conditions (Aichhorn et al., 2009; Döhnel et al., 2012;

Hooker et al., 2008; Sommer et al., 2010). On the one hand, false belief conditions may require more cognitive complexity or effort on the part of the subject. For that reason, one might hypothesize that the ToM cortical network should be more active in false belief trials than in true belief trials. On the other hand, the implied mental state of the agent in question is not necessarily different in false versus true belief trials. Thus, by modeling the same mental state, the ToM cortical network might respond in the same way to false and true belief trials. Comparing false and true belief trials, therefore, may be a less incisive test of the ToM network than comparing two different mental states attributed to an agent. The present study provides strong support to the contention that the right TPJ processes the inferred cognitive states of others.

Data availability

All data used in this study are available at https://figshare.com/s/f83f184793f8be13f37f.

Funding

This work was supported by the Princeton Neuroscience Institute Innovation Fund and Princeton Program in Cognitive Science. Arvid Guterstam was supported by the Wenner-Gren Foundation, the Swedish Brain Foundation, and the Promobilia Foundation.

Acknowledgements

We thank Jemma Bio-Barrick and Argos Wilterson for their encouragement.

References

Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G. 2014. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics* 8:14.

Aichhorn M, Perner J, Weiss B, Kronbichler M, Staffen W, Ladurner G. 2009. Temporo-parietal junction activity in theory-of-mind tasks: Falseness, beliefs, or attention. *Journal of Cognitive Neuroscience* 21:1179-1192.

Avants BB, Epstein CL, Grossman M, Gee JC. 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis* 12:26-41.

Baron-Cohen S. 1997. *Mindblindness: An Essay on Autism and Theory of Mind.* Cambridge (MA): MIT Press.

Baron-Cohen S, Leslie AM, Frith U. 1985. Does the autistic child have a 'theory of mind?' *Cognition* 21:37-46.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57:289-300.

Bio BJ, Webb TW, Graziano MSA. 2018. Projecting one's own spatial bias onto others during a theory-of-mind task. *Proceedings of the National Academy of Sciences USA* 115:E1684-E1689.

Brainard DH. 1997. The psychophysics toolbox. *Spatial Vision* 10:433-436.

Cox RW. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research* 29:162-173.

Döhnel K, Schuwerk T, Meinhardt J, Sodian B, Hajak G, Sommer M. 2012. Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *NeuroImage* 60:1652-1661.

Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, Oya H, Ghosh SS, Wright J, Durnez J, Poldrack RA, Gorgolewski KJ. 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods* 16:111-116.

Fletcher PC, Happé F, Frith U, Baker SC, Dolan RJ, Frackowiak RSJ, Frith CD. 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition* 57:109-128.

Fonov VS, Evans AC, McKinstry RC, Almli CR, Collins DL. 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* 47:S102.

Friston KJ, Holmes AP, Worsley KJ, Poline J-P, Frith CD, Frackowiak RS. 1994. Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping* 2:189-210.

Frith U, Frith CD. 2003. Development and Neurophysiology of Mentalizing. *Philosophical Transactions of the Royal Society of London, B Biological Sciences* 358:459-473.

Frith CD, Frith U. 2006. The neural basis of mentalizing. *Neuron* 50:531-534.

Gallagher HL, Frith CD. 2003. Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences* 7:77-83.

Gallagher HL, Happé F, Brunswick N, Fletcher PC, Frith U, Frith CD. 2000. Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia* 38:11-21.

Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, Haxby JV. 2007. Two takes on the social brain: a comparison of theory of mind tasks. *Journal of Cognitive Neuroscience* 19:1803-1814.

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS. 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics* 5:13 doi: 10.3389/fninf.2011.00013.

Greve DN, Fischl B. 2009. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* 48:63-72.

Grèzes J, Frith CD, Passingham RE. 2004. Inferring false beliefs from the actions of oneself and others: an fMRI study. *Neuroimage* 21:744-750.

Guterstam A, Bio BJ, Wilterson AI, Graziano MSA. (2021). Temporo-parietal cortex involved in modeling one's own and others' attention. *Elife* 10: e63551 doi: 10.7554/eLife.63551.

Hassabis D, Chu C, Rees G, Weiskopf N, Molyneux PD, Maguire EA. 2009. Decoding neuronal ensembles in the human hippocampus. *Current Biology* 19:546-554.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425-2430.

Hebart MN, Görgen K, Haynes J-D. 2015. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics* 8:88 doi: 10.3389/fninf.2014.00088.

Hooker CI, Verosky SC, Germine LT, Knight RT, D'Esposito M. 2008. Mentalizing about emotion and its relationship to empathy. *Social Cognitive and Affective Neuroscience* 3:204-217.

Igelström KM, Webb TW, Kelly YT, Graziano MSA. 2016. Topographical Organization of Attentional, Social, and Memory Processes in the Human Temporoparietal Cortex. *eNeuro* 3:ENEURO.0060-16.2016 doi: 10.1523/ENEURO.0060-16.2016.

Jenkinson M, Bannister P, Brady M, Smith S. 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* 17:825-841.

Kelly YT, Webb TW, Meier JD, Arcaro MJ, Graziano MSA. 2014. Attributing awareness to oneself and to others. *Proceedings of the National Academy of Sciences, USA* 111:5012-5017.

Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, USA* 103:38630-3868.

Lee J, Quintana J, Nori P, Green MF. 2011. Theory of mind in schizophrenia: exploring neural mechanisms of belief attribution. *Social Neuroscience* 6:569-581.

Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A. 2001. Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412:150-157.

Mar RA. 2011. The neural bases of social cognition and story comprehension. *Annual Reviews in Psychology* 62:103-134.

Marjoram D, Job DE, Whalley HC, Gountouna VE, McIntosh AM, Simonotto E, Cunningham-Owens D, Johnstone EC, Lawrie S. 2006. A visual joke fMRI investigation into Theory of Mind and enhanced risk of schizophrenia. *Neuroimage* 31:1850-1858.

Mitchell JP. 2009. Inferences about mental states. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences* 364:1309-1316.

Nakagawa S, Cuthill IC. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews* 82:591-605.

Poldrack RA. 2007. Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience* 2:67-70.

Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R. 2018. Development of the social brain from age three to twelve years. *Nature Communications* 9:1027 doi:10.1038/s41467-018-03399-2.

Rothmayr C, Sodian B, Hajak G, Döhnel K, Meinhardt J, Sommer M. 2011. Common and distinct neural networks for false-belief reasoning and inhibitory control. *NeuroImage* 56:1705-1713.

Saxe R. 2006. Uniquely human social cognition. *Current Opinion in Neurobiology* 16:235-239.

Saxe R, Kanwisher N. 2003. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind." *NeuroImage* 19:1835-1842.

Schurz M, Radua J, Aichhorn M, Richlan F, Perner J. 2014. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews* 42:9-34.

Sommer M, Meinhardt J, Eichenmüller K, Sodian B, Döhnel K, Hajak G. 2010. Modulation of the cortical false belief network during development. *Brain Research* 1354:123-131.

Spreng RN, Mar RA, Kim AS. 2009. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of Cognitive Neuroscience* 21:489-510.

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC. 2010. N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging* 29:1310.

van Veluw SJ, Chance SA. 2014. Differentiating between self and others: an ALE meta-analysis of fMRI studies of self-recognition and theory of mind. Brain Imaging and Behavior 8:24-38.

Vogeley K, Bussfeld P, Newen A, Herrmann S, Happé F, Falkai P, Maier W, Shah NJ, Fink GR, Zilles K. 2001. Mind Reading: Neural Mechanisms of Theory of Mind and Self-Perspective. *NeuroImage* 14:170-181.

Wellman HM. 2018. Theory of mind: The state of the art. *European Journal of Developmental Psychology* 15:728-755.

Wimmer H, Perner J. 1983. Beliefs About Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition* 13:103-128.

Zhang Y, Brady M, Smith S. 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging 20:45-57.

	L TPJ	R TPJ	L STS	R STS	MPFC	Precuneus
Decoding accuracy	49.8%	52.8%	49.4%	50.2%	49.7%	50.4%
95% CI	47.7 to 52.1	50.9 to 54.6	47.7 to 50.9	48.2 to 52.0	48.4 to 51.6	48.6 to 52.1
P value	0.518	0.004*	0.725	0.354	0.631	0.3290

Table 1. Decoding trials in which the cartoon "saw" a switch versus trials in which the cartoon "saw" no switch (nBS versus nBnS). For definition of ROIs, see Figure 2. Mean decoding accuracy (%), 95% confidence interval (based on bootstrap distribution), and p value (based on permutation testing, uncorrected for multiple comparisons) are shown for each of the six ROIs. The * indicates significant p values that survived correction for multiple comparisons across all six ROIs (FDR-corrected $p < 0.05$).

	L TPJ	R TPJ	L STS	R STS	MPFC	Precuneus
Decoding accuracy	48.6%	50.6%	50.1%	49.7%	49.4%	51.4%
95% CI	47.5 to 50.3	49.0 to 52.3	48.4 to 51.6	47.9 to 51.8	47.3 to 51.3	49.8 to 52.9
P value	0.936	0.260	0.420	0.683	0.776	0.0530

Table 2. Decoding false belief versus true belief trials (BS versus BnS). For definition of ROIs, see Figure 2. Mean decoding accuracy (%), 95% confidence interval (based on bootstrap distribution), and p value (based on permutation testing, uncorrected for multiple comparisons) are shown for each of the six ROIs. The * indicates significant p values that survived correction for multiple comparisons across all six ROIs (FDR-corrected $p < 0.05$).

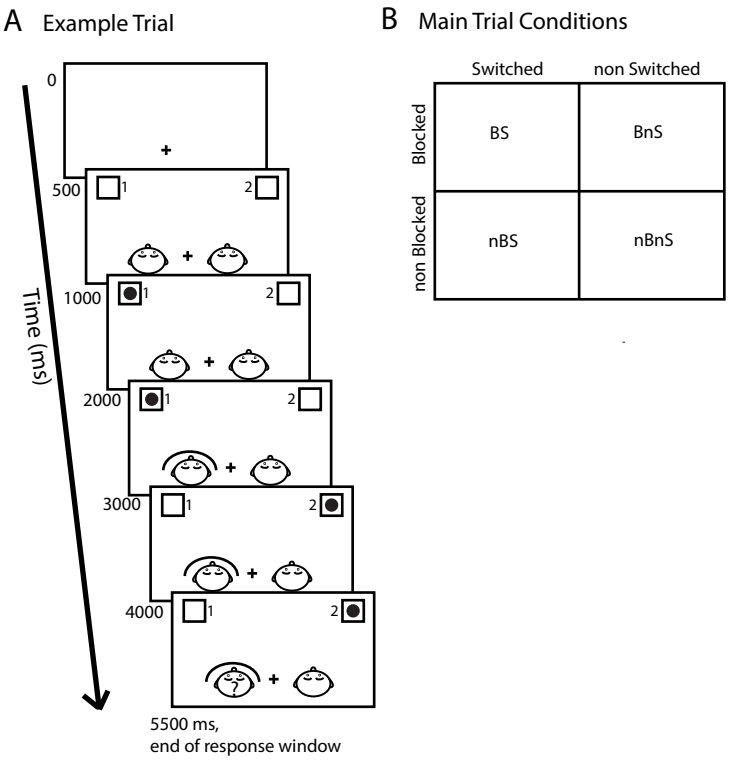

Figure 1. Behavioral paradigm. **A.** Timeline of events during a typical trial. Fixation point appeared at start of trial. Then two heads and two boxes appeared. Then a ball (colored red in the original stimulus) appeared in one box. Then one head had its sight blocked by the curved barricade. Then, on half of trials, the ball switched to the opposite box. Then a question mark appeared in one head, signaling subjects to respond by deciding whether the indicated head "thinks" the ball is in box 1 or box 2. All events were right-left counterbalanced among trials. **B.** Four main trial conditions formed by the 2 X 2 design of blocked versus nonblocked configurations and switched versus nonswitched configurations, resulting in blocked-switched (BS), blocked-nonswitched (BnS), nonblocked-switched (nBS) and nonblocked-nonswitched (nBnS) trials.

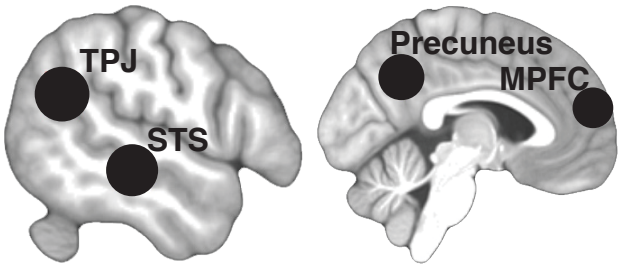

Figure 2. Regions of interest (ROIs). Six ROIs were defined based on peaks reported in an activation likelihood estimation meta-analysis of 16 fMRI studies involving theory-of-mind reasoning (van Veluw and Chance 2014). The ROIs consisted of 10-mm-radius spheres centered on peaks in the bilateral temporoparietal junction (TPJ) and superior temporal sulcus (STS), and two midline structures: the precuneus and medial prefrontal cortex (MPFC). Here, the TPJ and STS on the left side are shown. See **Materials and Methods** for ROI coordinates.

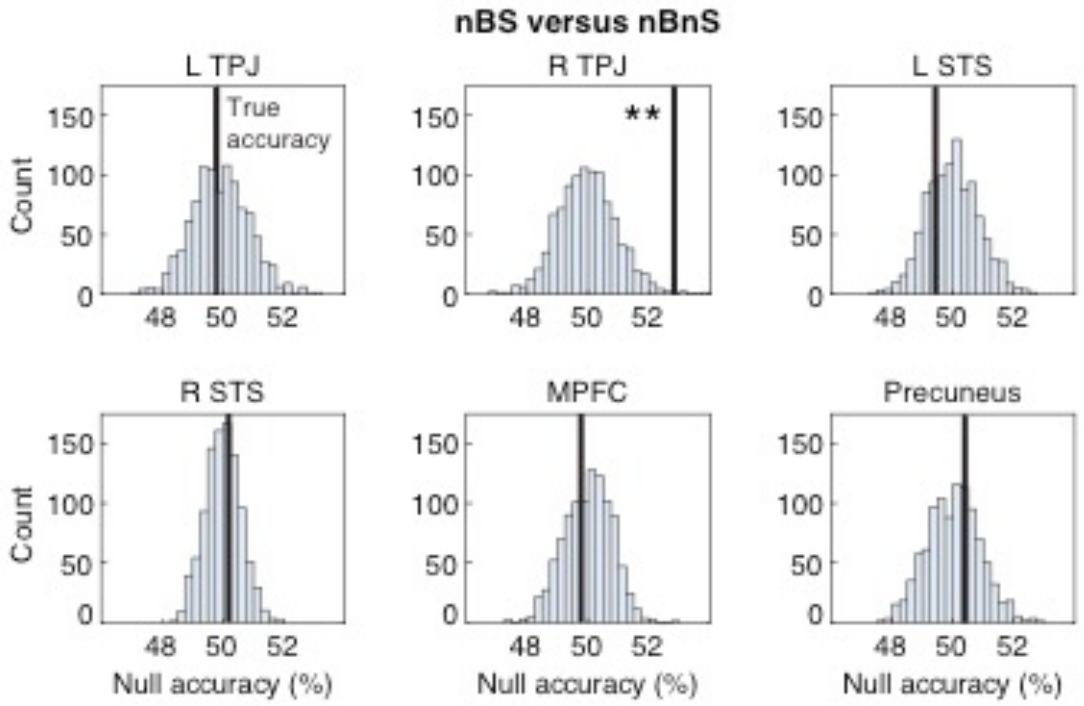

Figure 3. Decoding trials in which the cartoon "saw" a switch occur versus trials in which the cartoon "saw" that no switch occurred. Trials in which the cartoon "saw" a switch were represented by the nBS condition. Trials in which the cartoon "saw" no switch were represented by the nBnS condition. For definition of the six ROIs, see Figure 2. Each panel shows the results for one ROI. In each panel, the histogram shows the null distribution of decoding accuracies based on permutation testing with shuffled conditions labels (chance level = 50%). The tall vertical line placed within each histogram shows the accuracy of the classifier when it was trained and tested using the real (unshuffled) conditions labels. A decoding accuracy significantly greater than chance is indicated by * ($p<0.05$), based on permutation testing. The right TPJ showed significant decoding (p uncorrected = 0.004, p corrected using FDR for six ROIs = 0.024).

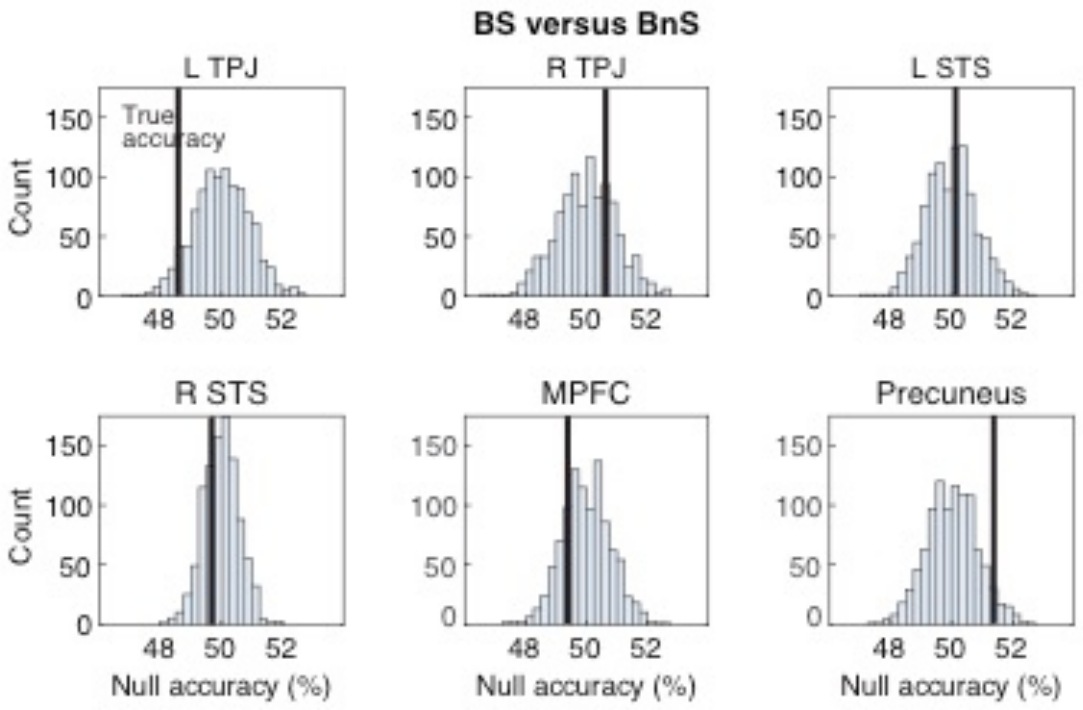

Figure 4. Decoding false belief versus true belief trials. Trials involving false belief were represented by the BS condition. Well-matched control trials involving true belief were represented by the BnS condition. For definition of the six ROIs, see Figure 2. Each panel shows the results for one ROI. In each panel, the histogram shows the null distribution of decoding accuracies based on permutation testing with shuffled conditions labels (chance level = 50%). The tall vertical line placed within each histogram shows the accuracy of the classifier when it was trained and tested using the real (unshuffled) conditions labels. Significance threshold ($p<0.05$) based on permutation testing, corrected for multiple comparisons across six ROIs using FDR. None of the ROIs showed significant decoding that distinguished false from true belief trials.