Right temporoparietal junction encodes inferred visual knowledge of others

Branden J. Bio a,1, Arvid Guterstam a,b,c,1, Mark Pinsk d, Andrew I. Wilterson a, Michael S.A. Graziano a,d,*

a Department of Psychology, Princeton University, Princeton, NJ, 08544, USA
b Department of Clinical Neuroscience, Karolinska Institute, 171 77, Solna, Stockholm, Sweden
c Department of Neurology, Karolinska University Hospital, Stockholm, Sweden
d Princeton Neuroscience Institute, Princeton University, Princeton, NJ, 08544, USA

ARTICLE INFO

Keywords:
False belief
fMRI
Mentalizing
Temporoparietal junction
Theory of mind

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.

1. 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., 2006; 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 Fig. 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. In many of the prior ToM studies cited above, it is impossible to determine exactly when, during an extended trial, the subject engages in a specific social cognition decision. In contrast, in a reaction-time task, it is possible to isolate a narrow time window in which a specific ToM decision occurs, and to examine brain activity during that time window.

We conceptualized the paradigm as a $2 \times 2$ design, as shown in Fig. 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, blocked-nonswitched, nonblocked-switched, and nonblocked-nonswitched.

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.

1.1. 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: nonblocked-switched and nonblocked-nonswitched. In these two conditions, the cartoon can see the ball at all times. In one trial type (nonblocked-nonswitched), it can see the ball placed in one box, and see that...
it stays there. In the other trial type (nonblocked-switched), 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 blocked-switched and the blocked-nonswitched 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 blocked-switched and the blocked-nonswitched trial types are the same. The cartoon head has no way to distinguish the two 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 nonblocked-nonswitched and nonblocked-switched trials. Second, the nonblocked-nonswitched versus nonblocked-switched contrast should be significantly greater than the blocked-switched versus blocked-nonswitched contrast. The blocked-switched versus blocked-nonswitched 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 nonblocked-nonswitched versus nonblocked-switched contrast, and the blocked-switched versus blocked-nonswitched contrast, the subject can see the ball switch boxes. But only in the nonblocked-nonswitched versus nonblocked-switched 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 such. We 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 et 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 response 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.

1.2. 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 sub-regions 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 blocked-switched and the blocked-nonswitched 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 (blocked-nonswitched) the head should have a true belief about the ball’s location, and when the ball is switched (blocked-switched) 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 nonblocked-switched and the nonblocked-nonswitched 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 blocked-nonswitched and blocked-switched trials. Second, the blocked-nonswitched versus blocked-switched contrast should be significantly greater than the nonblocked-switched versus nonblocked-nonswitched contrast. Hypothesis 2 therefore predicts exactly the opposite pattern of results as hypothesis 1.

2. Materials and methods

2.1. 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.

2.2. 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 scanner. During scanning, subjects laid in a supine position on the MRI scanner.
head would most likely believe the ball to be in box 1, to the left, or box one of two boxes and the participant had to decide whether a cartoon USA) was strapped to the subjects
for the ToM judgment.

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.

2.3. Behavioral task

The task events are illustrated in Fig. 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° 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° to the left (box 1) and right (box 2) of the midline, and 9° 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 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 2×2×2 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 Fig. 1B). These conditions formed a 2×2 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 Fig. 1B, these four conditions were labeled as blocked-switched, blocked-nonswitched, nonblocked-switched, and nonblocked-nonswitched.

Participants performed 256 trials (64 per main condition), in 8 runs of 32 trials each. Each run took approximately 5.5 min 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.

2.4. fMRI data acquisition

Functional imaging data were collected using a 3 T MAGNETOM Skyra (Siemens Healthineers AG, Erlangen, Germany) scanner equipped with a 64-channel head/neck coil. Gradient-echo T2weighted echo-planar images (EPI) with blood-oxygen dependent (BOLD) contrast was 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 inter-slice gap), which ensured that the entire brain excluding the cerebellum was within the field-of-view in all subjects (80 × 80 matrix, 2.5 mm × 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 vol 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 MPRAE 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.

2.5. 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:SCMR_005202) 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 ANTsv2.1.0 (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.readthed
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 Fig. 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 (4000 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 regresors), 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 $1/\text{total number of permutations}$, where the $1000$ trials are generated by flipping a coin, with a chance of $0.5$ of getting a particular outcome each time.

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.
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.

2.7. 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 family-wise error rate correction as implemented by SPM12.

3. Results

3.1. Task performance

The crucial importance of using a reaction-time task in the present experiment was to provide a narrow time window within which to target the MRI analysis. In most prior brain imaging experiments involving ToM, one does not know when, during an extended block of time (e.g. a 10 s trial), subjects may have been engaging in a specific social cognition judgement. But with the reaction time task, we knew that the decision process occurred at a specific time, within the allowed 1.5 s window. We therefore could target the MRI analysis much more precisely to that window of time, and hope to obtain more sensitive MRI measurements of hypothesized cognitive processes. Thus specific reaction-time data collected here is less important than the fact that a reaction-time task was used. The most important aspects of behavioral performance to check here are, first, whether subjects performed at high accuracy indicating proficiency with the task, and second, whether subjects responded within the 1.5 s response window. Subjects did perform the task at high levels of accuracy (overall 94.8% accuracy), suggesting that they understood the instructions and attributed beliefs to the cartoon heads as intended. Subjects responded within the allowed 1.5 s window in almost all trials (98%), and latency to respond was generally close to 1 s (overall latency 1006 ms). Table 1 shows the accuracy and latency data for all four conditions.

In addition to this basic assessment of behavioral performance, we also performed two specific, statistical comparisons on the behavioral data. In the MRI analysis described in the next section, two hypothesis-driven comparisons were made: first between the nonblocked-nonswitched condition and the blocked-switched condition, and second between the blocked-nonswitched condition and the blocked-switched condition. We therefore also tested the behavioral data for the same two comparisons, using targeted, paired t tests. For the contrast between the nonblocked-nonswitched condition and the blocked-switched condition, we found no significant difference in accuracy ($t(27) = -0.2$, $p = 0.829$), and a small significant difference in latency ($t(27) = 2.1$, $p = 0.045$). For the contrast between the blocked-nonswitched condition and the blocked-switched condition, we found a significant difference in accuracy ($t(27) = -3.3$, $p = 0.003$), and a significant difference in latency ($t(27) = 2.5$, $p = 0.020$). Although we did not have any specific a priori hypotheses concerning the behavioral data, these specific results are considered in relation to the MRI data in the discussion section.

3.2. MRI analysis on ROIs

According to hypothesis 1, cortical areas in the ToM network, especially the right TPJ, should show significant decoding for the nonblocked-nonswitched versus nonblocked-switched contrast. The results do show a significant decoding for this contrast in the right TPJ. Fig. 3 and Table 2 show the results for the ROI analysis, for the nonblocked-nonswitched versus nonblocked-switched contrast. In each panel, the thick vertical black line shows the accuracy of the MVPA decoding in which of the two trial types 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).

For the right TPJ (Fig. 3, top row, middle panel), the magnitude of decoding accuracy was 53%, compared to the chance level of 50%, and was highly statistically reliable ($p = 0.004$). Even when corrected for multiple comparisons across the six defined ROIs, the contrast remained statistically significant ($p = 0.024$ corrected using FDR). When subjects thought that the head could see the ball switch boxes, the right TPJ was affected by the switch, indicating that information about the switch-versus-nonswitch distinction was encoded in the activity within the right TPJ.

To further assess this result, we computed a confidence interval around the 53% accuracy result using a bootstrap distribution (see Methods). The decoder’s accuracy exceeded chance by more than the 95% confidence interval (see Table 2). A corrected $p$ value <$0.05$ in combination with a 95% confidence interval that does not cross chance level is standardly interpreted as a significant decoding effect at the group level (Nakagawa and Cuthill, 2007). In addition, we computed a Cohen’s effect size of $d = 0.56$, which is considered to be a medium-large effect. Finally, we performed a power analysis. In the comparison between the nonblocked-nonswitched condition and the nonblocked-switched condition, with 28 subjects and a decoding strength of 53%, the estimated power was 91%. A power of 80% or higher is generally considered to be the desired benchmark. Given these analyses, we can infer with a high degree of statistical confidence that the brain activity in the right TPJ contained information that distinguished between switched and nonswitched trial types, when subjects thought that the head could see the ball switch boxes.

According to hypothesis 2, cortical areas in the ToM network should show significant decoding for the blocked-nonswitched versus blocked-switched contrast. Fig. 4 and Table 3 show the results. None of the six ROIs showed any significant decoding (the decoding accuracy was not significantly different from the chance level of 50%; see Table 3 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. Hypothesis 2 was not supported.

Finally, we found that the decoding in the right TPJ for the nonblocked-nonswitched versus nonblocked-switched contrast was significantly greater than the decoding in the right TPJ for the blocked-

| Condition | BS | BnS | nBS | nBnS | Overall |
|-----------|----|-----|-----|------|---------|
| Accuracy (%) | Mean  | 91.6 | 95.2 | 96.0 | 96.2 | 94.8 |
| Latency (ms) | Mean | 1.4 | 1.0 | 0.8 | 0.9 | 0.9 |
| SEM | 1007 | 977 | 1035 | 1005 | 1006 |
| SEM | 29 | 31 | 22 | 28 | 26 |
3.3. MRI 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 nonblocked-switched and nonblocked-nonswitched trials revealed no significant areas of decoding at the brain-wide level; likewise, the searchlight comparison between blocked-nonswitched and blocked-switched trials revealed no significant areas of decoding at the brain-wide level.

3.4. 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 blocked-switched versus blocked-nonswitched and the nonblocked-switched versus nonblocked-nonswitched comparisons (i.e., blocked-switched > blocked-nonswitched, blocked-nonswitched > blocked-switched, nonblocked-switched > nonblocked-nonswitched, and nonblocked-nonswitched > blocked-switched). 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.

4. 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 switch ball 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 (nonblocked-nonswitched versus nonblocked-switched), not on trials when the subject could see the switch and the cartoon head could not (blocked-nonswitched versus blocked-switched). 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.

The results appear to be extremely subtle (53% decoding accuracy in the right TPJ for the nonblocked-nonswitched versus nonblocked-switched comparison, 3% greater than chance). However, it is easy to misinterpret the MVPA result. The MVPA signal is reported as a percentage that indicates the success rate of the machine classifier in distinguishing condition A from condition B. Here, with a small decoding success rate that is statistically reliable, we can make two inferences. First, information that distinguishes condition A from B is highly likely to be present in the brain activity in the right TPJ. Second, the information is noisy, such that it provides only a slight advantage to the decoder. For example, if every neuron in the TPJ responded perfectly to condition A but not to B—if the TPJ did nothing but encode that one signal without noise—then the decoder might achieve a 90% or higher decoding accuracy (a very unlikely situation almost never obtained in MVPA studies). In contrast, if a small percentage of neurons in the TPJ carry that information, while other neurons are active in ways that add noise to the signal, then the decoder might have a success rate only slightly above 50% (a more common situation). Yet a statistically reliable result would still indicate the presence of the relevant information.
in the brain activity in the TPJ. Here we have detected an information signal in the TPJ, and can infer its presence with a high degree of statistical confidence. Information about the distinction between the nonblocked-nonswitched and the nonblocked-switched conditions is evidently present in the TPJ activity, as predicted by hypothesis 1.

A possible alternative explanation of the MRI result is that, rather than reflecting cognitive decision processes, it might reflect a simpler difference in effort or task difficulty. For example, if subjects were significantly more accurate in one condition than another, then the activity in the TPJ might be related to greater or lesser task difficulty. However, the results of the behavioral analysis contradict this alternative hypothesis. The behavioral metrics (accuracy and latency) were significantly different between the two conditions that showed no decodable difference in the MRI signal (the blocked-nonswitched versus the blocked-switched conditions). Yet the behavioral measures were not significantly different (accuracy) or only marginally significantly different (latency) between the two conditions that did show a decodable difference in the MRI signal (the nonblocked-nonswitched versus the nonblocked-switched conditions). This result renders it extremely difficult to explain the MRI results as a product of slight performance differences between trial types.

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; Dohlen 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 for the contention that the right TPJ processes the inferred cognitive states of others.

The present results may also relate to the literature on perspective taking (e.g. Apperly and Butterfill, 2009; Kessler and Thomson, 2010; Martin et al., 2019). The MRI activation reported here is consistent with work using brain stimulation and perspective taking paradigms, that Martin et al., 2019. The MRI activation reported here is consistent with the right TPJ may play a complex, integrative role in a huge range of mental state, the ToM cortical network might respond in the same way to the right TPJ may play a complex, integrative role in a huge range of 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 for the contention that the right TPJ processes the inferred cognitive states of others.

The present results may also relate to the literature on perspective taking (e.g. Apperly and Butterfill, 2009; Kessler and Thomson, 2010; Martin et al., 2019). The MRI activation reported here is consistent with work using brain stimulation and perspective taking paradigms, that suggest a possible causal role for the right TPJ in understanding what others are thinking. 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 for the contention that the right TPJ processes the inferred cognitive states of others.

The present results may also relate to the literature on perspective taking (e.g. Apperly and Butterfill, 2009; Kessler and Thomson, 2010; Martin et al., 2019). The MRI activation reported here is consistent with work using brain stimulation and perspective taking paradigms, that suggest a possible causal role for the right TPJ in understanding what others are thinking. 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 for the contention that the right TPJ processes the inferred cognitive states of others.

The present results may also relate to the literature on perspective taking (e.g. Apperly and Butterfill, 2009; Kessler and Thomson, 2010; Martin et al., 2019). The MRI activation reported here is consistent with work using brain stimulation and perspective taking paradigms, that suggest a possible causal role for the right TPJ in understanding what others are thinking. 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 for the contention that the right TPJ processes the inferred cognitive states of others.

The present results may also relate to the literature on perspective taking (e.g. Apperly and Butterfill, 2009; Kessler and Thomson, 2010; Martin et al., 2019). The MRI activation reported here is consistent with work using brain stimulation and perspective taking paradigms, that suggest a possible causal role for the right TPJ in understanding what others are thinking. 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 for the contention that the right TPJ processes the inferred cognitive states of others.

The present results may also relate to the literature on perspective taking (e.g. Apperly and Butterfill, 2009; Kessler and Thomson, 2010; Martin et al., 2019). The MRI activation reported here is consistent with work using brain stimulation and perspective taking paradigms, that suggest a possible causal role for the right TPJ in understanding what others are thinking. 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 for the contention that the right TPJ processes the inferred cognitive states of others.

The present results may also relate to the literature on perspective taking (e.g. Apperly and Butterfill, 2009; Kessler and Thomson, 2010; Martin et al., 2019). The MRI activation reported here is consistent with work using brain stimulation and perspective taking paradigms, that suggest a possible causal role for the right TPJ in understanding what others are thinking. 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 for the contention that the right TPJ processes the inferred cognitive states of others.

The present results may also relate to the literature on perspective taking (e.g. Apperly and Butterfill, 2009; Kessler and Thomson, 2010; Martin et al., 2019). The MRI activation reported here is consistent with work using brain stimulation and perspective taking paradigms, that suggest a possible causal role for the right TPJ in understanding what others are thinking. 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 for the contention that the right TPJ processes the inferred cognitive states of others.

The present results may also relate to the literature on perspective taking (e.g. Apperly and Butterfill, 2009; Kessler and Thomson, 2010; Martin et al., 2019). The MRI activation reported here is consistent with work using brain stimulation and perspective taking paradigms, that suggest a possible causal role for the right TPJ in understanding what others are thinking. 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 for the contention that the right TPJ processes the inferred cognitive states of others.
