Processing of false belief passages during natural story comprehension: an fMRI study

Katerina D. Kandylaki[a,b], Arne Nagels[a], Sarah Tune[c], Richard Wiese[b], Ina Bornkessel-Schlesewsky[d], Tilo Kircher[a]

[a] Department of Psychiatry and Psychotherapy, Philipps-University Marburg, Rudolf-Bultmann-Str. 8, 35039 Marburg, Germany

[b] Department of Germanic Linguistics, Philipps-University Marburg, Deutschhausstr. 3, 35032 Marburg, Germany

[c] Department of Neurology, University of California, Irvine, Biological Sciences III, Irvine, CA 92697, USA

[d] Cognitive Neuroscience Laboratory, School of Psychology, Social Work and Social Policy, University of South Australia, Magill Campus, Adelaide, Australia.

Short title: False belief processing in auditory stories

Keywords: Theory of Mind | false beliefs | fMRI | auditory stories

Corresponding author: Katerina D. Kandylaki, Phone: +49(0)6421 58 63011, Address: Rudolf-Bultmann-Str. 8, 35039 Marburg, Germany, Email: kandylak@med.uni-marburg.de

Abstract

The neural correlates of theory of mind (ToM) are typically studied using paradigms which require participants to draw explicit, task-related inferences (e.g. in the false belief task). In a natural setup, such as listening to stories, false belief mentalising occurs incidentally as part of narrative processing. In our experiment, participants listened to auditorily presented stories with false belief passages (implicit false belief processing) and immediately after each story answered comprehension questions (explicit false belief processing), while neural responses were measured with fMRI. All stories included (amongst other situations) one false belief condition and one closely matched control condition. For the implicit ToM processing, we modelled the hemodynamic response during the false belief passages in the story and compared it to the hemodynamic response during the closely matched control passages. For implicit mentalising we found activation in typical ToM processing regions, i.e. the angular gyrus (AG), superior medial frontal gyrus (SmFG), precuneus (PCUN), middle temporal gyrus (MTG) as well as in the inferior frontal gyrus (IFG) billaterally. For explicit ToM we only found AG activation. The conjunction analysis highlighted the left

AG and MTG as well as the bilateral IFG as overlapping ToM processing regions for both implicit and explicit modes. Implicit ToM processing during listening to false belief passages , recruits the left superior medial frontal gyrus and billateral precuneus in addition to the ”mentalising network” known form explicit processing tasks.

Introduction

Theory of Mind (ToM) is the cognitive capacity to attribute mental states to self and others (Goldman, Margolis, Samuels, & Stich, 2012; Premack & Woodruff, 1978). This capacity, which is also referred to as mentalising (e.g. C. Frith & Frith, 1999; U. Frith & Frith, 2010), is employed incidentally in a broad range of naturally occurring social situations. Consider, for example, a situation in which two friends are talking and a third joins the discussion after two minutes. The first two will be aware that the third will have no knowledge of their preceding conversation: they have a “Theory of Mind” (Premack & Woodruff, 1978) that allows them to take on the third person’s perspective. Although the term ToM may

appear relatively straightforward from the perspective of this initial definition, the rich facet of social situations where ToM is employed remains to be systematically categorised (Schaafsma, Pfaff, Spunt, & Adolphs, 2014). ToM is thought to involve a variety of sub- and super-processes (Schaafsma et al., 2014) and can be deconstructed based on criteria such as implicit versus explicit and cognitive versus affective (e.g. Schlaffke et al., 2014). In addition to the deconstruction of the concept of ToM, and in order to quantify ToM as precisely as possible, a reconstruction of ToM components from basic building blocks is needed: for example, face recognition and gaze processing are essential for completing the reading the mind in the eyes test (RMET), which is a measure of mentalising ability (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). FMRI research on specific ToM tasks, such as false belief processing, can inform this reconstruction with brain maps capturing instances of ToM, for details see Schaafsma et al. (2014). In spite of these multifaceted aspects of ToM, several meta-analyses of functional neuroimaging studies connected to ToM processing (Mar, 2011; Northoff et al., 2006; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014; Van Overwalle, 2009) suggest that a number of regions appear to be

involved in ToM regardless of the specific task. These include the bilateral temporo-parietal junction (TPJ), which corresponds to the inferior portion of the angular and supramarginal gyri (AG, SMG) as well the posterior superior temporal sulcus (pSTS), superior parts of the frontal gyrus (SFG), medial prefrontal cortex (mPFC) and precuneus (PCUN).

Van Overwalle and colleagues first investigated another aspect of spontaneous ToM processing in trait attributions (Kestemont, Vandekerckhove, Ma, Van Hoeck, & Van Overwalle, 2013; Ma et al., 2012; Ma, Vandekerckhove, Van Overwalle, Seurinck, & Fias, 2011) and causal attributions (Kestemont et al., 2014); in these studies, spontaneous and intentional ToM processing was manipulated between participants via the experimental instructions. In another study, *implicit* mentalising was elicited via visually presented images in the contrast of false vs. true belief processing (Sommer, Döhnel, Sodian, Meinhardt, Thoermer, & Hajak, 2007). According to a recent meta-analysis (Schurz et al., 2014), trait attribution and false belief processing show very similar activation patterns. This seems meaningful conceptually because causal attributions refer to beliefs about a temporary event just like false beliefs. However, stimulus

presentation was very similar in many of the studies for trait attribution and false belief processing that were included in the meta-analysis: all stimuli were visual and included a minimal amount of contextual information.

Here, by contrast, we present the first study to employ auditory story stimuli in conjunction with a rich situational context, within which the false belief situations were embedded. We assume that ToM processing in such embedded false belief passages is *implicit* because it occurs during listening to a story with the simple goal of keeping track of the narrative plot. If the story involves situations that require the attribution of mental states to others, mentalising would be a prerequisite for successful understanding of the plot. This notion is similar to van der Wel, Sebanz, & Knoblich's (2014) automatic belief tracking, although they measured behaviour in the movement trajectory of a cursor while the subject was giving their response, after having watched a short movie which involved belief tracking. By contrast, a situation or task in which a participant was asked to actively answer questions or make judgements about a person's mental state can be described as *explicit* ToM with regard to belief processing. In this case, the inferences drawn regarding another person's mental state are clearly task-

induced.

False belief processing in story comprehension

Story-based approaches as opposed to highly controlled experimental setups provide the opportunity of testing false belief processing in the context of a semantically rich, cohesive story.

The first story-based neuroimaging study on ToM was reported by Fletcher et al. (1995). The authors presented short stories each involving one false belief or control situation and examined the contrast between false belief vs. physical stories vs. unlinked sentences. Participants were instructed to read the stories and answer one question immediately after each story. They were, however, informed about the type of the story, i.e., in the case of false belief stories, that they should pay attention to people's beliefs and intentions. Also, the authors modelled false belief processing during the reading of false belief and control passages and during the answering of a related question together. For the critical contrast between ToM and physical stories, this study observed activation in left medial frontal regions and in the anterior and posterior cingulate cortex (aCC,

pCC).

The majority of subsequent story-based neuroimaging investigation of ToM processing used the original false belief vignettes from the Fletcher et al. (1995) PET study or translations thereof and compared them to different control conditions. All of them presented the stories visually and modelled participants' reading process. For example, Vogeley et al. (2001) modified this material in order to cross ToM and self versus other processing. Story, question and silent answer were all modelled together in a block design; thus, there was no differentiation between *implicit* and *explicit* false belief processing. For the main effect of ToM, Vogeley and colleagues found differences in (among others): the right aCC, right superior frontal gyrus (SFG), and left lateral prefrontal cortex (lPFC).

Using fMRI, Saxe & Kanwisher (2003) contrasted false belief and mechanical inference stories in two reading experiments and found ToM-related activation in the TPJ both for reading a false belief story (Experiment 1) and for reading the story and performing the question answering task (Experiment 2). This region has shown robust activation across a number of story-based studies on ToM

processing (Aichhorn et al., 2009; Gallagher et al., 2000; Kobayashi, Glover, & Temple, 2006; Lee, Quintana, Nori, & Green, 2011; Saxe & Kanwisher, 2003; Spengler, Cramon, & Brass, 2009).

Aichhorn et al. (2009) tested the contrast of false belief vs. photograph during story processing and were the first to explicitly distinguish two different modes in ToM processing: Time point *1: Story*, when the participants read the story (again, *implicit* false belief processing in reading) and Time point *2: Question*, when the participants read and answered the question about the story (*explicit* false belief processing in reading). Their results showed activation in the middle and superior temporal gyri (MTG, STG), superior and inferior frontal gyri (SFG, IFG) and in the TPJ for false belief compared to false photograph conditions at both time points. Also, precuneus activation was found only for the task but not for the story.

In summary, the story-based paradigms which investigated *explicit* false belief processing have identified a network comprising the following regions: mPFC, SFG, IFG, TPJ, MTG, and precuneus. However, all previous story-based studies used a block design in which one story and one answer formed one condition, either the false belief or the control condition. Most of these studies modelled

implicit and *explicit* ToM together in the same block (Fletcher et al., 1995; Saxe & Kanwisher, 2003; Vogeley et al., 2001), with the notable exception of Aichhorn et al. (2009), who modelled them separately (but in reading comprehension and in short stories). One previous fMRI study did use auditory linguistic stimuli to contrast causality-related inferences with ToM-related inferences (Ferstl & Cramon, 2002). However, this experiment only employed minimal context and did not distinguish between ToM-related inferences drawn during listening and task performance.

The design details of previous studies (mostly visual stimulation with minimal context and an explicit task) may have contributed to some of the inconsistencies about the involvement of the TPJ in false belief processing that are apparent in the literature (Callejas et al. 2011). More specifically, most of the above mentioned studies did not differentiate between components such as text processing or memory maintenance. The current study, by contrast, extended the work of Ferstl & Cramon (2002) by using linguistically rich narratives and by embedding false belief passages into these larger contexts. This allowed us to model the processing of false belief passages in language comprehension

(implicit ToM) separately from overall story processing and decision-making during question answering. In comparison to previous designs, this manipulation should render results less sensitive, for example, to individual processing speed (as in reading paradigms). In addition, we aimed to capture ToM-related processes when reading and answering a question about false beliefs. To this end, we modelled reading of both question and answer options during which the participants manipulated the story information in order to perform the task (explicit ToM).

Ferstl & Cramon (2002) raised the question of the relationship between coherence processes in narrative comprehension and ToM processing. For both logical (coherence) and person-related (ToM) inferences, the results mainly shared frontomedian cortex (FMC) activation. The authors connect this activation to a domain general function of the FMC, “the initiation and maintenance of non automatic cognitive processes” (Ferstl & Cramon, 2002, p. 1610). This view accords well with a recent review on ToM, which argues that ToM-related processes should be decomposed into smaller subprocessing blocks (Schaafsma et al. 2014). From this perspective, we could assume that both

language and ToM are high-level functions that comprise a number of basic subprocesses (for a recent neurobiological perspective that advocates the decomposition of language into basic submechanisms, see Bornkessel-Schlesewsky, Schlesewsky, Small, & Rauschecker, 2015). Some of the subprocesses might overlap: for example, one basic subprocess of ToM (according to Schaafsma et al., 2014) may be the understanding of causality, which is also a basic subprocess in the semantics of language comprehension (Kuperberg et al. 2011). It may be the case that, when these subprocesses overlap, brain activations in the respective networks are enhanced. This might be one possibility of how linguistic and social processes like ToM interact in the brain, although this interaction remains to be studied more systematically.

The present study

The present fMRI study tested *implicit* and *explicit* false belief processing by means of a novel paradigm: a) we presented stories (approximately 2 minutes in length) with false belief and control situations embedded amongst a range of other scenarios; b) these stories were presented auditorily; and c) this design

allowed us to model the hemodynamic response to false belief and control events based on their onset and duration within the two-minute-long story. We chose to have participants listen to the stories (instead of reading them) in order to increase naturalness in the setup, since reading is a culturally recent innovation that is less than 6000 years old (see also (Dehaene & Cohen, 2011)'s "neuronal recycling hypothesis"). In order to also examine *explicit* mentalising, we presented participants with two questions subsequent to each story, targeting both false belief as well as control story content. The questions and answers were presented visually, in order to provide participants with various information modalities, which would keep them alert throughout the whole experiment.

Based on previous research on story-based false belief processing, we expected to observe activation for the contrast of false belief ToM vs. control passages during story listening (*implicit* false belief processing) in the mentalising network: mPFC, SFG, IFG, MTG, precuneus and TPJ. Based on the Aichhorn et al. (2009) findings, we expected *explicit* false belief processing activation to manifest itself mainly in the rTPJ and SFG as part of the mPFC (see Schurz et al. (2014) for the role of the mPFC in mentalising).

Methods

Participants

Twenty-two monolingual native speakers of German participated in the study, all right-handed (Edinburgh Inventory of Handedness) (age mean = 24.3 years, sd = 2.1 years, male N = 6), recruited from postings at the University of Marburg. We had to exclude data from two participants due to movement artefacts, resulting in a total of 20 datasets that entered the final analysis for the current study. The study was approved by the ethics committee of the Faculty of Medicine of the University of Marburg. All participants gave written informed consent before participating in the study and were paid 30 euros for participation.

Stimuli

For testing *implicit* ToM processing we created 20 stories with a length of 2 minutes (±10 seconds; mean and standard deviation of story length 306 (13) words, 23 (4) sentences). All stories included one false belief condition (im-TOM) and one control condition (im-NONTOM). The 40 (im-TOM and im-NONTOM)

situations included a variety of social interactions: 12 every day situations e.g. *cooking, playing, driving*, 12 scenarios related to hobbies such as *hiking, visiting an art exhibition, sports*, 10 work situations in e.g. *shop, office, conference*, 4 school situations (*school trip, chess competition*) and 2 university situations (*department party*). The common pattern of all situations was that there were two persons involved and one of them had a false belief. In order to comprehend these passages the participants needed to use their theory of mind. The control passages were matched in length to the false belief passages and were part of a physical chain of events, in which one event led to the next one. The important difference between false belief and control passages was the existence of "different minds", the situation in which two story participants have different beliefs. In the analysis of the results, the even chain passages which did not require mentalising are referred to as control or NONTOM passages.

We developed two versions of each story as follows: version A included im-TOM at the end of the first minute of the story and im-NONTOM (control passage) at the end of the second minute of the story. In version B the manipulation was reversed: im-NONTOM (control passage) at the end of the first

minute and im-TOM at the end of the second minute of the story. This design resulted in a total of 40 stories (20×2 minimal pairs). To avoid one participant hearing two versions of the same story (with a similar plot but alternating critical passages), we split the 40 stories into two lists of 20 stories each. One participant heard one of the two lists in an individually randomised order: for example Participant 1 would hear list 1 which contained story 1A, 2B, 3B, 4A and so on, and Participant 2 would hear list 2 which contained story 1B, 2A, 3A, 4B and so on. Examples of the critical passages from two versions of the same story (1A and 1B) were as follows (translated from the German original):

Story 1A

- Within the first half: "[…] but his wife was so busy taking pictures of the idyllic landscape, so she didn't realise, that her husband ate all the sandwiches. When later they arrived at the summit of Brocken, she also wanted to eat a sandwich, but found only drinks in her bag. *She thought that maybe she had forgotten the food in the car.* […]"
  im-TOM

- Within the second half: "[…] The man took the camera from his backpack and

gave it to the hiker. The hiker put in a lot of effort and *took a whole series of pictures from all different perspectives.* (im-NONTOM) The couple thanked him and went on walking. […]"

Story 1B

- Within the first half: "[...] A few minutes later she came back to her husband and he had already eaten three sandwiches. At the summit of the Brocken the woman got also hungry and she found the last sandwich. *With delight she ate it and drank a few sips of apple juice.* (im-NONTOM) [...]"

- Within the second half: "[…] The man gave the camera to the hiker who accidentally packed it up in his backpack after photographing the couple. Then he went on hiking. *The couple was looking for the camera without success,* (im-TOM) until the hiker came back and apologised many times. [...]"

Stimuli were spoken by a professionally trained female speaker of German at a

normal speech rate. We recorded the stimuli in a sound proof EEG laboratory cabin with a sampling rate of 44.1 kHz and a 16bit (mono) sample size. For sampling we used the sound recording and analysis software Amadeus Pro (version1.5.3, HairerSoft) and an Electret microphone (Beyerdynamic MC930C). Two example stories are available in the supplementary material.

We pre-tested the stories prior to the imaging study in order to validate their quality. In an online questionnaire we gathered ratings from 177 participants. The questionnaire was distributed through a students' mailing list. Participants who did not fulfil the language criteria (monolingually raised German native speakers) were excluded from the final analysis. The participants were asked to judge comprehensibility (“How comprehensive was this passage?”) and naturalness (“How natural was this passage?”) of the auditory stimuli. Ratings were collected on a 4-point scale from 1 (very unnatural / incomprehensible) to 4 (extremely natural / comprehensible). The use of earphones was highly recommended in the instructions of the questionnaire. An analysis using linear mixed effects models using R statistical software (Team R. Core, 2014) and the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) with fixed factor of condition and random

effects (only intercepts due to convergence problems) of story and subject showed that ratings for im-TOM and im-NONTOM passages did not differ significantly: Comprehensibility means (standard deviations): im-TOM 3.62 (0.62) vs. im-NONTOM 3.64 (0.57), $p = 0.67$, Naturalness means (sds): im-TOM 3.23 (0.79) vs. im-NONTOM 3.22 (0.77), $p = 0.87$.

In the scanner, the stories were presented auditorily, while the subject was looking at a fixation point in the centre of a computer display. After each story, two questions and two possible answers for each question were presented visually. The questions referred to the im-TOM and im-NONTOM part of the story, thus creating the *explicit* conditions ex-TOM and ex-NONTOM. For example, the questions and answers for the above-mentioned example passages were:

Story 1A

- ex-TOM: Where did the woman think that the sandwiches were? Answers: In the car vs. in her husband's stomach
- ex-NONTOM: Who had the camera when there was a series of pictures taken? Answers: The hiker vs. the old man

Story 1B

- ex-NONTOM: How many sandwiches were left for the woman, after the man had finished eating? Answers: One vs. three
- ex-TOM: Where was the camera in the opinion of the couple, after the hiker took pictures of them? Answers: In the backpack of the couple vs. in the backpack of the hiker

In order to rule out possible alternative explanations of the results for the ex-TOM vs. ex-NONTOM contrast, questions and answers were analysed according to the metrics in Table 1. The following metrics were analysed using linear mixed effects models (package lme4) in R: 1. question length in words, 2. number of clauses of question (as a measurement of syntactic complexity), 3. answer length in words and 4. question type, whether it asked about location or not. In the inferential statistics we used likelihood ratio tests to compare: a) the null model, in which only the random factor of story is included and b) the main effect of ToM, in which the type of the question (ex-TOM vs. ex-NONTOM) as well as the random factor of story is included. For question and answer length the main effect of ToM model showed a marginally significant improvement over the null model (p = 0.0643 and p = 0.08183, respectively). For the criteria of number of clauses and question type (location vs. non-location) there was no significant

difference between the null and the main effect of ToM models (p = 0.1137 and p = 0.495, respectively).

Imaging procedure and behavioural data acquisition

Prior to the scanning procedure, a training session outside the scanner was performed. Participants listened to two stories and answered two questions subsequent to each story. The practice stimuli were not part of the experimental stimuli. In the scanner the participants listened to 20 stories and answered 40 questions (two after each story). Participants heard the stories through MRI compatible earphones. Sound quality and loudness was optimised in the scanner before starting the experiment. The order of the stories was assigned randomly and was different for each participant, in order to avoid sequence effects. The stories were divided into 4 blocks of 5 stories each. After each block the participant had a break of 45 seconds. During the break the participants saw the visual message “Short break!” in the middle of the screen, while the scanner was still running.

One story trial consisted of the following events: first a fixation cross was shown in the middle of the screen for 500 ms before the story started. The cross

was then replaced by a fixation point and at the same time the story started. The duration of the story was approximately 2 minutes. After the story there was a jitter between 1.5 and 4.5 seconds (duration assigned randomly), after which the first question was presented visually. The question was presented all at once, centred and towards the top of the screen for 5 seconds. After that, the possible answers appeared towards the bottom of the screen, clearly separated from each other; each answer began with an index letter a. always on the left, and b. always on the right side of the screen below the question (see Figure 1 for a graphical representation of the question and answer screens). The possible answers stayed on the screen until participants made their decision; however, they disappeared if participants took longer than 3 seconds to respond (duration pretested in order to ensure a natural pace of the experiment). Participants gave their answers by pressing the left or right button on a button box, which was fixed to their left leg, with their left middle or index finger accordingly. The left hand was chosen as a response hand in order to minimise left hemispheric artefacts which could overlap with linguistic processing (Callan et al. 2004). The position of the correct answer was counterbalanced. Presentation of stimuli was time-jittered

between story and questions and also between first and second question. All visual stimuli (cross, fixation point, questions and answers) were presented in dark grey on light grey background. Figure 1 shows a graphical representation of an example trial. The procedure was implemented and presented with the software package Presentation (Neurobehavioral Systems Inc., San Francisco, CA).

Behavioural data analyses

For the behavioural data analyses we used R statistical software (R Core Team, 2014) and the lme4 package (Bates, Maechler, Bolker, & Walker, 2013). For both responses and reaction times (RTs) we calculated models with fixed factors of condition and question order and random factors story and subject. The first question always referred to the first manipulation in the story (irrespective of the condition) and the second question always tested the information of the second manipulation of the story (also irrespective of the condition). We used logistic regression (because both dependent and independent variables were categorical) in combination with the maximal random effects structure (random

slopes and intercept per condition and answer order for story and subject; see Barr, Levy, Scheepers, & Tily, 2013) for the response analyses (R function glmer). Due to convergence problems in the models with the maximal random effects structure in the RTs, we included in the models (calculated with the R function lmer) the most complex random effects structure that reached convergence (random slopes and intercept per condition for story and subject). To assess the effects of the different factors on the response times and responses, we employed a forward model selection procedure within which we used likelihood ratio tests to compare a base model including only an intercept with successively more complex models (function anova in R).

fMRI data acquisition

During the MR-session a series of echo-planar-images was gathered to record the time course of the subjects' brain activity. Measurements were performed on a 3 Tesla MRI system (Trio, A Tim System 3T, Siemens, Erlangen, Germany) with a 12 channel head matrix receive coil. Functional images were acquired using a T2* weighted single shot echo planar imaging (EPI) sequence: parallel imaging

factor of 2 (GRAPPA), TE=25ms, TR=1450ms, flip angle 90°, slice thickness 4.0 mm and 0.6 mm gap, matrix 64×64, field of view = 224×224 mm, in-plane resolution 3.5×3.5mm2, bandwidth 2232Hz/pixel, EPI factor of 64 and an echo spacing of 0.53 ms. Transversal slices oriented to the AC–PC line were gathered in ascending order.

The initial five images were removed from the analyses in order to avoid saturation and stabilization effects. Head movements of the participants were minimised by using foam paddings.

A whole head T1 weighted data set was acquired with a 3d MPRage sequence (parallel imaging factor of 2 (GRAPPA), TE=2.26ms, TR=1900ms, flip angle 9°, 1 mm isometric resolution, 176 sagittal slices, 256×256 matrix).

fMRI data analyses

All analyses for the fMRI data were calculated in SPM8 (Welcome Trust Centre for Neuroimaging), implemented in MATLAB (Mathworks Inc., Sherborn, MA).

A slice time correction (to the 15th slice) was performed first. Then images were realigned to the first image in order to correct for head movement artefacts.

We normalised the volumes into standard stereotaxic anatomical Montreal Neurological Institute (MNI) space by using the transformation matrix calculated from the first EPI scan of each subject and the EPI template. On the normalised data (resliced voxel size 2mm3) we applied an 8 mm full-width-at-half-maximum (FWHM) Gaussian smoothing kernel in order to compensate for inter-subject anatomical variation.

For the single-subject analysis the design matrix for each subject was created individually, based on the log files from the fMRI-session, because each participant heard the stories in a different order. We modelled im-TOM and im-NONTOM conditions in seconds (mean duration of event = 4326 ms, standard deviation = 1525 ms). As critical events we modelled one sentence from each passage: for im-TOM it was the sentence in which the protagonist had a false belief and for im-NONTOM it was a length-matched sentence from the control passage. The events of the previous examples were as follows:

Story 1A

- *She thought that maybe she had forgotten the food in the car.*
  im-TOM

- *took a whole series of pictures from all different perspectives.*
im-NONTOM

Story 1B

- *With delight she ate it and drank a few sips of apple juice.*
im-NONTOM

- *The couple was looking for the camera without success,*
im-TOM

For ex-TOM and ex-NONTOM conditions we modelled the question (5s) and answer (RTs) trial together as critical events (mean duration: 7.426s, sd: 1.030s); these events did not involve the motor response. The two trials for which there was no response were not modelled. As factors of no interest we modelled separately: the rest of the stories (excluding the im-TOM and im-NONTOM parts), the button presses (motor responses) and the jitters before each question and story. Our baseline consisted of the three 45s pauses between blocks. In order to remove movement artefacts for each individual session the realignment parameters were entered as multiple regressors in the first-level analysis.

On the group-level analysis, we modelled two *T*-contrasts between the

following first-level conditions: a) im-TOM vs. im-NONTOM and b) ex-TOM vs. ex-NONTOM. Brain activations were plotted on the anatomical MRIcron (http://www.mccauslandcenter.sc.edu/mricro/mricron/) high resolution template (the Colin brain). We used the cluster extent thresholding algorithm by (Slotnick, Moo, Segal, & Hart, 2003), which employs a FWE correction using a Monte Carlo simulation approach, in order to correct for multiple comparisons. We set the desired correction threshold for multiple comparisons to $p < 0.05$ and the assumed voxel type I error to $p < 0.001$; after 10000 iterations our cluster extend threshold was estimated at 48 voxels. For all fMRI results reported for *implicit* (im-TOM vs. im-NONTOM) and *explicit* (ex-TOM vs. ex-NONTOM) contrasts (and the reverse contrasts), we employed a whole brain analysis and used an individual voxel threshold of $p < 0.001$ with a cluster extend threshold of 48 voxels.

Contrasts of interest

In addition to the contrasts im-TOM vs. im-NONTOM and ex-TOM vs. ex-NONTOM, which test *implicit* and *explicit* ToM separately, we were interested in the common regions activated for both contrasts. Therefore we performed a

conjunction analysis using statistical parametric maps (SPMs) of the minimum *T*-statistic over the previous contrasts (im-TOM vs. im-NONTOM and ex-TOM vs. ex-NONTOM). Inference was based on p-values adjusted for the search volume using random field theory (for details on the exact procedure see Friston, Penny, & Glaser, 2005). The SPM8 algorithm for conjunction (testing the conjunction null hypothesis as recommended in (Nichols, Brett, Andersson, Wager, & Poline, 2005)) assumes that the p-value of the conjunction is the square root of the p-value of the involved contrasts. We set the p-value for the conjunction to 0.05, thereby implicitly thresholding each individual contrast at $p < 0.0025$. We further corrected the conjunction results for multiple comparisons by setting a cluster threshold on 120 voxels, as estimated by the Slotnick et al. (2003) algorithm after 10000 iterations.

Results

Subjects achieved a mean of 90% (sd=5.61) correctness in the answers. The mean percentages of correct, incorrect and missed responses per condition are presented in Table 2. In a logistic mixed effects models analysis (see

"Behavioural data analyses" for details) we found no significant main effect of condition (ex-TOM vs. ex-NONTOM) for the responses (correct, incorrect and missed response). A comparison of the logistic mixed effects models (using the function anova in R) showed no significant improvement of model fit (*p's* > .7) of the main effects and interaction models in comparison to the null model (which included only an intercept in addition to the maximal random effects structure). None of the single main effects models (only main effect of condition or only main effect of question order) improved model fit compared to the null model (*p's* > 0.5).

We also analysed the response times (RTs) with mixed effects models (for the results see Figure 2). We found no significant improvement of the model fit (p = 0.26) when comparing the main effects model of condition and question order to the null model (which included only the intercept and random effects). None of the single main effects of condition or question order improved the model in comparison to both the null model and the main effects of condition and question order models (*p's* > 0.1).

In the fMRI analyses we found significant activation for im-TOM vs. im-

NONTOM in the bilateral angular gyrus (AG), left MTG, right middle temporal pole (MTP), bilateral precuneus (PCUN), left cerebellum (CE) (crus 2, VIII), right CE (IX), as well as bilateral MFG and IFG. For ex-TOM vs. ex-NONTOM, we found significant activation in the left AG. For the reverse implicit contrast im-NONTOM vs. im-TOM we found activation in the right MFG and left posterior central gyrus (PoCG) (p <.001, cluster extend threshold of 48 voxels). For the reverse explicit contrast ex-NONTOM vs. ex-TOM we found activation in the right hippocampus (HC) (also p <.001, cluster extend threshold of 48 voxels). An overview of the results for the contrasts of interest with coordinates (MNI), *T*-scores and cluster sizes is presented in Table 3 (see Figure 3 for the localisation of the results on the brain template).

Table 4 shows the supra-threshold clusters activated for the conjunction. We found supra-threshold activation for the conjunction in the left AG, MTG, MFG bilateral IFG as well as in the cerebellum (crus 1) (for the localisation of the effects see Figure 5).

The changes of the BOLD signal, as reflected in the contrast estimates (first principal component of the signal) in the peak voxel, in the left AG across all

contrasts of interest revealed more activation for the im-TOM and ex-TOM in comparison to the im-NONTOM and ex-NONTOM condition respectively (see Figure 4). In the barplots of the contrast estimates for the implicit contrast (see Figure 4 and Figure 6, under the red label), it is the differences between im-TOM and im-NONTOM that are responsible for the *implicit* ToM activation. In the ex-TOM vs. ex-NONTOM plots (see Figure 4, under the yellow label) the differences between ex-TOM and ex-NONTOM are driving the supra-threshold activation. Please note that in both implicit and explicit contrasts (ex-TOM vs. ex-NONTOM under the yellow label and im-TOM vs. im-NONTOM under the red label) we have plotted all conditions of interest for the sake of completeness, even though only two are relevant for the activation of the contrast. For the conjunction contrast (see Figure 4 and Figure 6 under the green label), all four conditions contribute to the common activation of implicit and explicit contrasts by showing the same tendencies in the differences between TOM and NONTOM conditions: a left lateralised network comprising the AG, MTG, MFG and IFG showed stronger increases in BOLD signal for im-TOM and ex-TOM in comparison to im-NONTOM and ex-NONTOM.

Discussion

In the present study, participants listened to two-minute long stories with embedded *implicit* false belief and control passages. After each story participants answered one false belief and one control question (*explicit* false belief processing). Our main finding was that processing false beliefs *incidentally* in a rich and natural narrative context recruits ToM processing regions (AG, MTG, mPFC in the SmFG, precuneus) – known from the *explicit* ToM literature. For *explicit* false belief processing we could replicate previous results in our left AG findings. The conjunction analysis revealed a left lateralised network of the AG, MTG, MFG and IFG as the common pattern activated during both *implicit* and *explicit* false belief processing.

For *implicit* false belief processing (im-TOM vs. im-NONTOM) we found activation in the AG bilaterally. The AG is an anatomical subdivision of the posterior inferior parietal lobule (pIPL), which is considered part of the so-called temporo-parietal junction (TPJ) (Mar, 2011; Seghier, 2013). TPJ activation has been previously found in several story-based ToM paradigms (Aichhorn et al.,

2009; Fletcher et al., 1995; Gallagher et al., 2000; Kobayashi et al., 2006; Lee et al., 2011; Mitchell, 2008; Saxe & Kanwisher, 2003; Vogeley et al., 2001). Especially the right TPJ is assumed to be heavily involved in mentalising processes (Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Saxe & Kanwisher, 2003) throughout many different situations: answering an open question about visually presented and cartoon stories (Gallagher et al., 2000), answering in a multiple choice question after visually presented stories (Spengler et al., 2009), second-order false belief tasks (Kobayashi et al., 2006), false belief stories and desires also presented visually (Saxe & Kanwisher, 2003), reading TOM and SELF stories and covertly answering questions (Vogeley et al., 2001). All of these previous studies did not investigate or model *implicit* false belief processing separately, although it was also measured during stimulation. The only study from the literature that attempted to capture *implicit* false belief processing was Aichhorn et al. (2009), in which the reading of false belief stories was modelled separately to answering the question in two different first level analyses. Although another instance of automatic ToM processing has been tested in trait judgements (Ma et al., 2012, 2011), the current study was the first

to test *implicit* ToM in false belief processing a) in listening and b) embedded in a rich and natural linguistic context. Since we found TPJ activation for *implicit* false belief processing in this novel setup, we conclude that the TPJ may not only be active in decision-based setups but that it may also be recruited incidentally in implicit false belief tasks (such as mentalising during narrative processing).

Our contribution adds to the existing literature on social cognition as part of language processing (Ferstl & Cramon, 2002; Ferstl, Neumann, Bogler, & Cramon, 2008; Mason & Just, 2009) by revealing ToM processing regions when people listen to stories including false belief situations. Our finding of SmFG (mPFC) in the implicit contrast and in the conjunction in particular is in line with Ferstl & Cramon (2002)'s findings of FMC activation for narrative coherence based on social cues. However, a more systematic study of the interplay between social and linguistic cues in narrative processing is needed in order to draw conclusions about how the two domains interact on a neurobiological level.

The TPJ is one of the regions of a ToM processing *network*. This network includes (amongst others) the following regions: mPFC, IFG, MFG, MTG, CE and PCUN (Aichhorn et al., 2009; Gallagher et al., 2000; Kobayashi et al., 2006; Lee

et al., 2011; Saxe & Kanwisher, 2003; Spengler et al., 2009). In the present study we provided further support for this ToM network especially during *implicit* false belief processing when listening to narrative stories. Our findings specifically provide evidence for the claim that ToM processing regions are recruited spontaneously when false beliefs are embedded in natural context.

Parts of the “classical” ToM processing network have been claimed to be strongly involved in the default mode network (DMN). Especially the AG and the precuneus have repeatedly shown reliable activation during resting state fMRI experiments (Shehzad et al., 2009; Seghier, 2013; Utevsky et al. 2014). In connection to our findings and given our experimental setup, our results strengthen the claims for these areas (AG, PCUN) to be involved in (automatic) belief tracking, as explained by van der Wel et al. (2014). A similar concept to the DMN is the human "reorienting" system (Corbetta, Patel, & Schulman, 2008), which distinguishes between a dorsal and a ventral attention system. Even though these systems are formulated in connection to visual cognition we would like to attempt a connection to auditory processing. In this view the AG and the precuneus belong to the dorsal attention system and support top-down

attentional control. On the other hand, the ventral attention network includes the MFG, the ventral frontal cortex (VFC), and the inferior parts of the TPJ (posterior STG) and is responsible for bottom-up reorientation. This framework could offer an alternative interpretation for our results in terms of attentional reorienting (for example as in Rothmayr et al., 2011).

Our *explicit* false belief contrast showed supra-threshold activation only in the left AG. Since there has been only one study which aimed to disentangle implicit and explicit false belief processing (Aichhorn et al., 2009) we can attempt a comparison of our explicit contrast results with their findings at *Time Point 2: Question.* Their main contrast was false belief (FB) vs. false photograph (PH). As this contrast was used in order to define the ROIs for the remaining contrasts, a fact that highlights the importance of this contrast in comparison to the remaining contrasts of interest, we can compare it to our explicit contrast (ex-TOM vs. ex-NONTOM). Their results also showed activation in the left AG, as in our manipulation, but they included additional areas such as the anterior MTG, temporal pole, IFG, PCUN and mPFC. These were precisely the areas that showed signal changes for our implicit contrast, while they did not survive the

threshold for the explicit contrast. However, the conjunction pointed to the same clusters (left lateralised AG, MTG, IFG, MFG, CE) as commonly activated areas for implicit and explicit contrasts. This suggested that the explicit contrast mask extended to the other ToM processing areas but the effect sizes were not strong enough to survive the $p < .001$ threshold. Moreover, methodological details of design and modelling might also explain the differences in the findings of our explicit contrast to the FB vs. PH contrast of Aichhorn et al. (2009). The design differed in four respects: the presentation modality (listening vs. reading stories); the length of the stories and therefore the amount of provided context (23 sentences vs. 2 sentences); the first level analysis (modelled critical events in one first level analysis vs. modelled critical events in two different first level analyses); the second level analysis (whole-brain vs. combined whole-brain and ROI analysis). Finally, due to the absence of jitters (else, the presence of a constant ISI of 2s) between the modelled events in Aichhorn et al. (2009) it is unclear how reliably the contributions of each conditions to the overlapping BOLD response could have been estimated. In contrast to this, in our study of narrative stories, the context of the story created a natural jitter between the events of the

implicit contrast. Also, for the explicit contrast we optimised our design for efficient modelling of the hemodynamic response by introducing ISIs (jitters) of random duration (1.5 – 4.5s) between story and question as well as between the first and second question.

Despite the single cluster activation for the explicit contrast, the pattern of left lateralised activation revealed by the conjunction of the two contrasts of interest was in line with the ToM processing network known from the imaging literature. This left lateralisation might be connected to the nature of the stimuli, which comprised of false belief situations presented auditorily and embedded in rich linguistic context.

We have to acknowledge two minor limitations in our paradigm. First, it might be the case that the questions (*explicit* ToM) pointed participants towards what was tested in the study. However, these questions comprised only half of the total number of asked questions. There was always a control question in addition to the false belief question after each story and the order of the two questions was counterbalanced across the whole experiment. Moreover, the formulations of the false belief questions were very variable, so that we did not repeat the words

“think”, “opinion”, etc., too often during the experiment. In addition, the stories include a wide variety of situations of which the experimental manipulations constitute not even half of the total duration of one story. Thus, participants can be keeping track of more things at different times during the story: e.g. they might have followed the path that the hikers took (in the hiking story mentioned in the Stimuli section).

The second limitation is related to the order of the presented conditions; in our design the *implicit* condition always preceded the *explicit* condition, which means that the two conditions were not totally independent in time. This might have caused more common activation than if these two conditions were measured completely independently. However, other experimental solutions, such as asking the false belief questions after the fMRI scan session, would also be suboptimal, since remembering the plots of 20 stories would be very demanding and lead to a high number of incorrect answers. We chose to ask two questions after each story not only in order to test false belief processing in *explicit* mode, but also to make sure that the participants were alert and paying attention to the stories. Since two of our participants almost fell asleep repeatedly during scanning (these

data sets were excluded from the analyses), we have to accept that it might be difficult to lay in the MRI scanner for one hour listening to stories without falling asleep, despite the noisy environment of the scanning procedure and the task.

Conclusions

In the present study we showed that the "classical" ToM processing network (e.g. TPJ, mPFC, MTG, PCUN) is activated during *implicit* false belief processing, while listening to short stories. We were the first to reliably disentangle *implicit* and *explicit* ToM processing by modelling short false-belief sentences within longer stories and by separating them futher from the explicit false belief task. Our study therefore offers insights to the neural underpinnings of auditory language processing and social competence, integral parts of human nature long before the use of written communication.

Acknowledgments

We would like to thank Karen Henrich and Marie-Josephine Rocholl for their help in recording the stories as well as Jens Sommer, Mechthild Wallnig and Rita Werner for their help in data collection. This project was funded by the ExInit initiative of Philipps-University Marburg.

References

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*(6), 1179–92. doi:10.1162/jocn.2009.21082

Apperly, I., & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. doi:10.1037/a0016923

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 42(2), 241–51.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-

effects models using Eigen and S4. *R package version*, *1*(4).

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S.L., & Rauschecker, J.P. (2015). Neurobiological roots of language in primate audition: common computational properties. *Trends in Cognitive Sciences, 19,* 142-150.

Callan, D. E., Jones, J. A., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native-and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. *NeuroImage*, *22*(3), 1182-1194.

Callejas, A., Shulman, G. L., & Corbetta, M. (2011). False belief vs. false photographs: a test of theory of mind or working memory?. *Frontiers in psychology*, *2*.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, *58*(3), 306–24. doi:10.1016/j.neuron.2008.04.017

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–62. doi:10.1016/j.tics.2011.04.003

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*(3), 1652–1661. doi:10.1016/j.neuroimage.2012.01.073

Ferstl, E. C., & Cramon, D. Y. von. (2002). What does the frontomedian cortex contribute to language processing: coherence or theory of mind? *NeuroImage*, *17*, 1599–1612. doi:10.1006/nimg.2002.1247

Ferstl, E. C., Neumann, J., Bogler, C., & Cramon, D. Y. von. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, *29*, 581–593. doi:10.1002/hbm.20422

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of “theory of mind” in story comprehension. *Cognition*, *57*(2), 109–28. Retrieved

from http://www.ncbi.nlm.nih.gov/pubmed/8556839

Friston, K. J., Penny, W. D., & Glaser, D. E. (2005). Conjunction revisited. *NeuroImage*, *25*(3), 661–7. doi:10.1016/j.neuroimage.2005.01.013

Frith, C., & Frith, U. (1999). Interacting minds – a biological basis. *Science*, *286*(5445), 1692–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10576727 http://www.sciencemag.org/content/286/5445/1692.short

Frith, U., & Frith, C. (2010). The social brain: allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1537), 165–76. doi:10.1098/rstb.2009.0160

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, *38*(1), 11–21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10617288

Goldman, A. I., Margolis, E., Samuels, R., & Stich, S. (2012). *Theory of Mind*.

Hartwright, C. E., Apperly, I., & Hansen, P. C. (2012). Multiple roles for executive control in belief-desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. *NeuroImage*, *61*(4), 921–930. doi:10.1016/j.neuroimage.2012.03.012

Kestemont, J., Ma, N., Baetens, K., Clément, N., Van Overwalle, F., & Vandekerckhove, M. (2014). Neural correlates of attributing causes to the self, another person and the situation. *Social Cognitive and Affective Neuroscience*, 1–8. doi:10.1093/scan/nsu030

Kestemont, J., Vandekerckhove, M., Ma, N., Van Hoeck, N., & Van Overwalle, F. (2013). Situation and person attributions under spontaneous and intentional instructions: An fMRI study. *Social Cognitive and Affective Neuroscience*, *8*, 481–493. doi:10.1093/scan/nss022

Kobayashi, C., Glover, G. H., & Temple, E. (2006). Cultural and linguistic influence on neural bases of 'Theory of Mind': an fMRI study with Japanese bilinguals. *Brain and Language*, *98*(2), 210–20. doi:10.1016/j.bandl.2006.04.013

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: a neural prediction problem.

Neuron, *79*(5), 836–48. doi:10.1016/j.neuron.2013.08.020

Kuperberg, G. R., Paczynski, M., & Ditman, T. (2011). Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience*, *23*(5), 1230-1246.

Lee, J., Quintana, J., Nori, P., & Green, M. F. (2011). Theory of mind in schizophrenia: exploring neural mechanisms of belief attribution. *Social neuroscience*, *6*(5-6), 569-581.

Luyten, P., & Fonagy, P. (2014). Mentalising in attachment contexts. *The Routledge Handbook of Attachment: Theory*, 107.

Ma, N., Vandekerckhove, M., Baetens, K., Overwalle, F. V., Seurinck, R., & Fias, W. (2012). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience*, *7*, 937–950. doi:10.1093/scan/nsr064

Ma, N., Vandekerckhove, M., Van Overwalle, F., Seurinck, R., & Fias, W. (2011). Spontaneous and intentional trait inferences recruit a common mentalising network to a different degree: spontaneous inferences activate only its core areas. *Social Neuroscience*, *6*(2), 123–38. doi:10.1080/17470919.2010.485884

Mar, R. a. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, *62*, 103–34. doi:10.1146/annurev-psych-120709-145406

Mars, R. B., Sallet, J., Schüffelgen, U., Jbabdi, S., Toni, I., & Rushworth, M. F. S. (2012). Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. *Cerebral Cortex (New York, N.Y. : 1991)*, *22*(8), 1894–903. doi:10.1093/cercor/bhr268

Mason, R. a., & Just, M. A. (2009). The role of the theory-of-mind cortical network in the comprehension of narratives. *Linguistics and Language Compass*, *3*, 157–174. doi:10.1111/j.1749-818X.2008.00122.x

Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex (New York, N.Y. : 1991)*, *18*(2), 262–71. doi:10.1093/cercor/bhm051

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, *25*(3), 653–60.

doi:10.1016/j.neuroimage.2004.12.005

Northoff, G., Heinzel, A., Greck, M. de, Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain–a meta-analysis of imaging studies on the self. *NeuroImage*, *31*(1), 440–57. doi:10.1016/j.neuroimage.2005.12.002

Packard, M. G., & Knowlton, B. J. (2002). Learning and memory functions of the Basal Ganglia. *Annual Review of Neuroscience*, *25*, 563–93. doi:10.1146/annurev.neuro.25.112701.142937

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: the roles of left and right temporo-parietal junction. *Social Neuroscience*, *1*(3-4), 245–58. doi:10.1080/17470910600989896

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioural *and Brain Sciences*, *1*(04), 515–526.

Rilling, J. K., Sanfey, A. G., Aronson, J. a, Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions.

NeuroImage, *22*(4), 1694–703. doi:10.1016/j.neuroimage.2004.04.015

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*(3), 1705-1713.

Sallet, J., Mars, R. B., Noonan, M. P., Neubert, F.-X., Jbabdi, S., O'Reilly, J. X., … Rushworth, M. F. (2013). The organization of dorsal frontal cortex in humans and macaques. *The Journal of Neuroscience: the Official Journal of the Society for Neuroscience*, *33*(30), 12255–74. doi:10.1523/JNEUROSCI.5108-12.2013

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*(4), 1835–1842. doi:10.1016/S1053-8119(03)00230-1

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2014). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, 1–8. doi:10.1016/j.tics.2014.11.007

Schlaffke, L., Lissek, S., Lenz, M., Juckel, G., Schultz, T., Tegenthoff, M., … Brüne, M. (2014). Shared and non shared neural networks of cognitive and

affective theory-of-mind: A neuroimaging study using cartoon picture stories. *Human Brain Mapping*, *00*(August). doi:10.1002/hbm.22610

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*, *42C*, 9–34. doi:10.1016/j.neubiorev.2014.01.009

Seghier, M. L. (2013). The angular gyrus multiple functions and multiple subdivisions. *The Neuroscientist*, *19*(1), 43-61.

Shehzad, Z., Kelly, A. C., Reiss, P. T., Gee, D. G., Gotimer, K., Uddin, L. Q., ... & Milham, M. P. (2009). The resting brain: unconstrained yet reliable. *Cerebral cortex*, *19*(10), 2209-2229.

Slotnick, S. D., Moo, L. R., Segal, J. B., & Hart, J. (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. *Brain Research. Cognitive Brain Research*, *17*(1), 75–82. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12763194

Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G.

(2007). Neural correlates of true and false belief reasoning. *Neuroimage*, *35*(3), 1378-1384.

Spengler, S., Cramon, D. Y. von, & Brass, M. (2009). Control of shared representations relies on key processes involved in mental state attribution. *Human Brain Mapping*, *30*(11), 3704–18. doi:10.1002/hbm.20800

Spreng, R. N., Mar, R. a, & Kim, A. S. N. (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*(3), 489–510. doi:10.1162/jocn.2008.21029

Team, R. C. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012.

Utevsky, A. V., Smith, D. V., & Huettel, S. A. (2014). Precuneus is a functional core of the default-mode network. *The Journal of Neuroscience*, *34*(3), 932-940.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*, 829–858. doi:10.1002/hbm.20547

Vogeley, K., Bussfeld, P., Newen, a, Herrmann, S., Happé, F., Falkai, P., …

Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *NeuroImage*, *14*(1 Pt 1), 170–181. doi:10.1006/nimg.2001.0789

Wel, R. P. R. D. van der, Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*(1), 128–33. doi:10.1016/j.cognition.2013.10.004

Legends

Figure 1: One example trial. The trial began with a fixation cross, then a fixation point was shown during the whole story (approximately 2 minutes). Next, the question screen appeared for 5 seconds and after this the answer screen (which still included the question) was shown for maximally 3 seconds. After that, the screens of second question and answer appeared with the same durations as the first question and answer screens.

Figure 2: Reaction times (RTs) for the First and Second question after each story for the ex-TOM and ex-NONTOM conditions. Abbreviations: cor: correct response, inc: incorrect response.

Figure 3: The contrast im-TOM vs. im-NONTOM is depicted in red and ex-TOM vs. ex-NONTOM in yellow, superimposed on a high resolution brain template (the Colin brain) of the MRIcron software (p <.001, cluster extend threshold of 48 voxels, Monte Carlo corrected). Abbreviations: AG: angular gyrus, SmFG: superior medial frontal gyrus, mPFC: medial prefrontal cortex, MTG: middle temporal gyrus, MFG: middle frontal gyrus, CE: cerebellum, PCUN: precuneus.

Coordinates: a) MNI: -56, -58, 32, b) MNI: -60, -24, -10, c) MNI: 46, 22, 38, d) MNI: -8, -48, 44.

Figure 4: Bar plots of BOLD signal changes (contrast estimates) at the peak voxel of the left (LH) angular gyrus (AG) for implicit and explicit false belief processing as well as their conjunction. Threshold for the implicit and explicit contrasts: p <.001 and cluster extend threshold of 48 voxels (Monte Carlo corrected), for the conjunction p <.05 and cluster extend threshold of 120 voxels (Monte Carlo corrected). The error bars represent 90% confidence intervals. Please note that the peak voxel of the LH AG is the same for the explicit contrast and the conjunction, therefore the barplots are also the same.

Figure 5: The activations of the conjunction of the contrasts im-TOM vs. im-NONTOM and ex-TOM vs. ex-NONTOM are shown in green (conjunction voxel threshold of *p* <.05 and cluster extend threshold of 120 voxels, Monte Carlo corrected). Abbreviations: AG: angular gyrus, MTG: middle temporal gyrus, MFG: middle frontal gyrus, IFG: inferior frontal gyrus. Coordinates: a) MNI: -56, -58, 32, b) MNI: -52, -46, -14, c) MNI: -42, 28, -12.

Figure 6: Bar plots of BOLD signal changes (contrast estimates) at the peak voxel

of each cluster for implicit false belief processing and the conjunction of implicit and explicit contrasts. Abbreviations: RH: right hemisphere, LH: left hemisphere, AG: angular gyrus, SmFG: superior medial frontal gyrus, mPFC: medial prefrontal cortex, MTG: middle temporal gyrus, MTP: middle temporal pole, MFG: middle frontal gyrus, PCUN: precuneus. Please note that we have placed the plots for precuneus, AG and IFG next to each other for illustration reasons and with absolutely no intention of denoting that they are overlapping regions. For explicit and implicit contrasts: individual voxel threshold of $p < .001$ and cluster extend threshold of 48 voxels (Monte Carlo corrected). For the conjunction: individual voxel threshold of $p < .05$ and cluster extend threshold of 120 voxels (Monte Carlo corrected). The error bars represent 90% confidence intervals.

Table 1: Question and answer details for the contrast ex-TOM vs. ex-NONTOM. The inferential statistics represent model comparison of two models: a) the null model, in which only the random factor of story is included and b) the main effect of ToM, in which the type of the question (ex-TOM vs. ex-NONTOM) as well as the random factor of story is included.

Metric	All	ex-TOM	ex-NONTOM
	Mean (standard deviation)		
Question length in words	8.9 (3)	9.5 (3.2)	8.3 (2.8)
		p = 0.06	
Number of clauses of the question	1.5 (0.5)	1.6 (0.5)	1.4 (0.5)
		p = 0.11	
Answer length in words	3.7 (1.9)	3.9 (2.2)	3.5 (1.5)
		p = 0.08	
	Percentage (Number)		
Location content of question	58.8% (47/80)	62.5% (25/40)	55% (22/40)
		p = 0.49	

Table 2: Mean percentage of correct, incorrect and missed responses per condition.

Answer	ex-TOM	ex-NONTOM
Correct	92	88
Incorrect	7	11
Not answered	1	2

Table 3: False belief activation peaks with their local maxima coordinates for the contrasts im-TOM vs. im-NONTOM and ex-TOM vs. ex-NONTOM (p <.001, cluster extend threshold 48

voxels). Coordinates are listed in MNI atlas space (H: hemisphere, mPFC: medial prefrontal cortex).

Contrast	Anatomical region	H	MNI Coordinates	*T*	Cluster size
im-TOM vs. im-NONTOM	Angular gyrus (AG)	L	-58 -64 30	7.64	2121
	Cerebellum (CE) – Crus 2	L	-24 -82 -34	7.19	2853
	Angular gyrus (AG)	R	56 -60 30	7.1	2812
	Precuneus (PCUN)	L	-8 -48 44	6.23	860
	Middle frontal gyrus (MFG)	L	-22 52 24	5.89	1620
	Superior medial frontal gyrus (SmFG) – mPFC	L	-4 48 38	5.68	1620
	Middle frontal gyrus (MFG)	L	-40 12 46	5.28	806
	Middle temporal pole (MTP)	R	50 8 -28	5.03	257
	Middle frontal gyrus (MFG)	R	24 24 44	4.97	236
	Middle temporal gyrus (MTG)	L	-60 -24 -10	4.71	410
	Precuneus (PCUN)	R	12 -50 40	4.4	860
	Superior frontal gyrus (SFG)	R	26 58 18	4.4	88
	Middle frontal gyrus (MFG)	R	46 22 38	4.3	217
	Superior frontal gyrus (SFG)	L	-32 52 0	4.3	52
	Inferior frontal gyrus (IFG) – Pars triangularis	L	-56 26 6	4.11	297
	Inferior frontal gyrus (IFG) – Pars orbitalis	L	-50 28 -6	4	297
ex-TOM vs. ex-NONTOM	Angular gyrus (AG)	L	-56 -58 32	3.76	75

Table 4: Activation peaks with their local maxima coordinates for the conjunction of the contrasts im-TOM vs. im-NONTOM and ex-TOM vs. ex-NONTOM (threshold $p < .05$, cluster minimum of

120 voxels, Monte Carlo corrected). Coordinates are listed in MNI atlas space (H: hemisphere).

Contrast	Anatomical region	H	MNI Coordinates	T	Cluster size
Conjunction	Angular gyrus (AG)	L	-56 -58 32	3.76	1363
	Middle frontal gyrus (MFG)	L	-40 20 48	2.61	487
	Middle temporal gyrus (MTG)	L	-52 -36 -14	2.42	484
	Inferior frontal gyrus (IFG) – Pars triangularis	R	54 30 28	2.37	135
	Cerebellum (Crus 1)	R	28 -80 -32	2.33	293
	Inferior frontal gyrus (IFG)	L	-42 28 -12	2.31	139

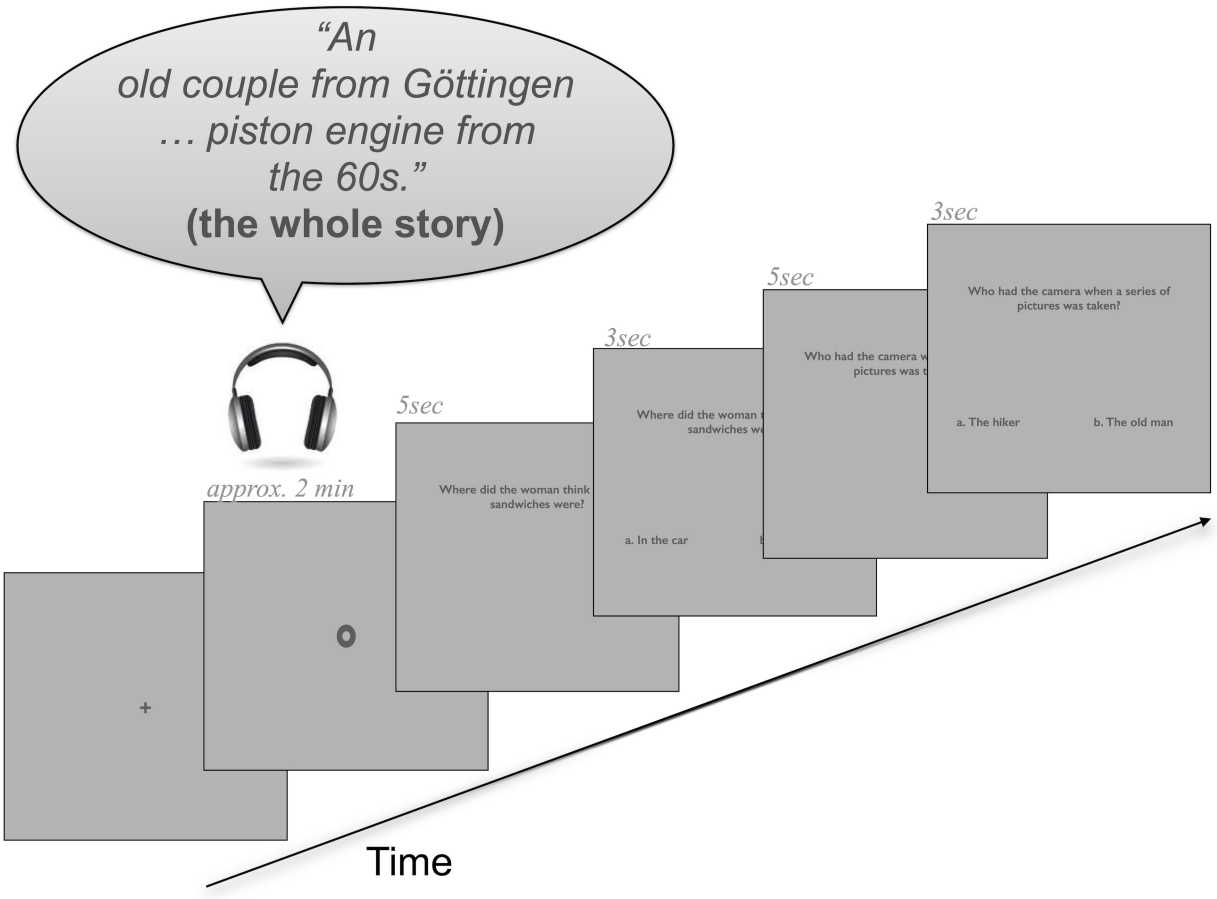

Figure 1: One example trial. The trial began with a fixation cross, then a fixation point was shown during the whole story (approximately 2 minutes). Next, the question screen appeared for 5 seconds and after this the answer screen (which still included the question) was shown for maximally 3 seconds. After that, the screens of second question and answer appeared with the same durations as the first question and answer screens.

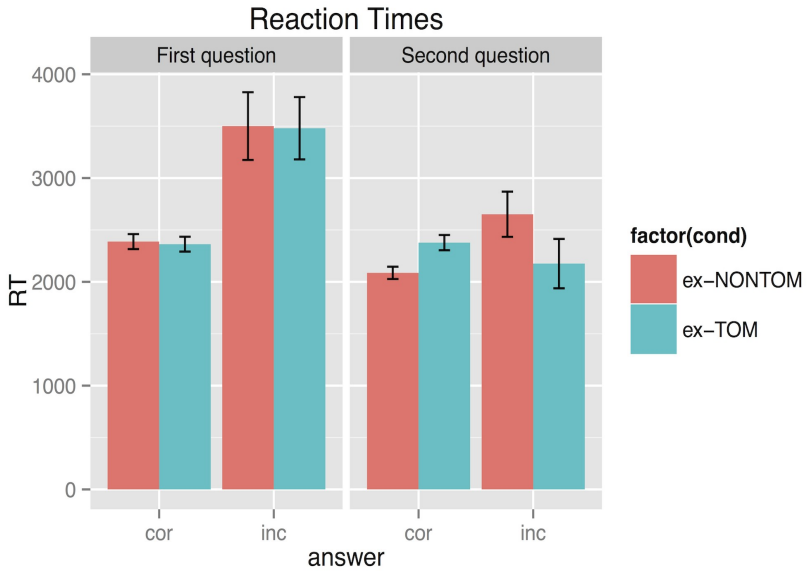

Figure 2: Reaction times (RTs) for the First and Second question after each story for the ex-TOM and ex-NONTOM conditions. Abbreviations: cor: correct response, inc: incorrect response.

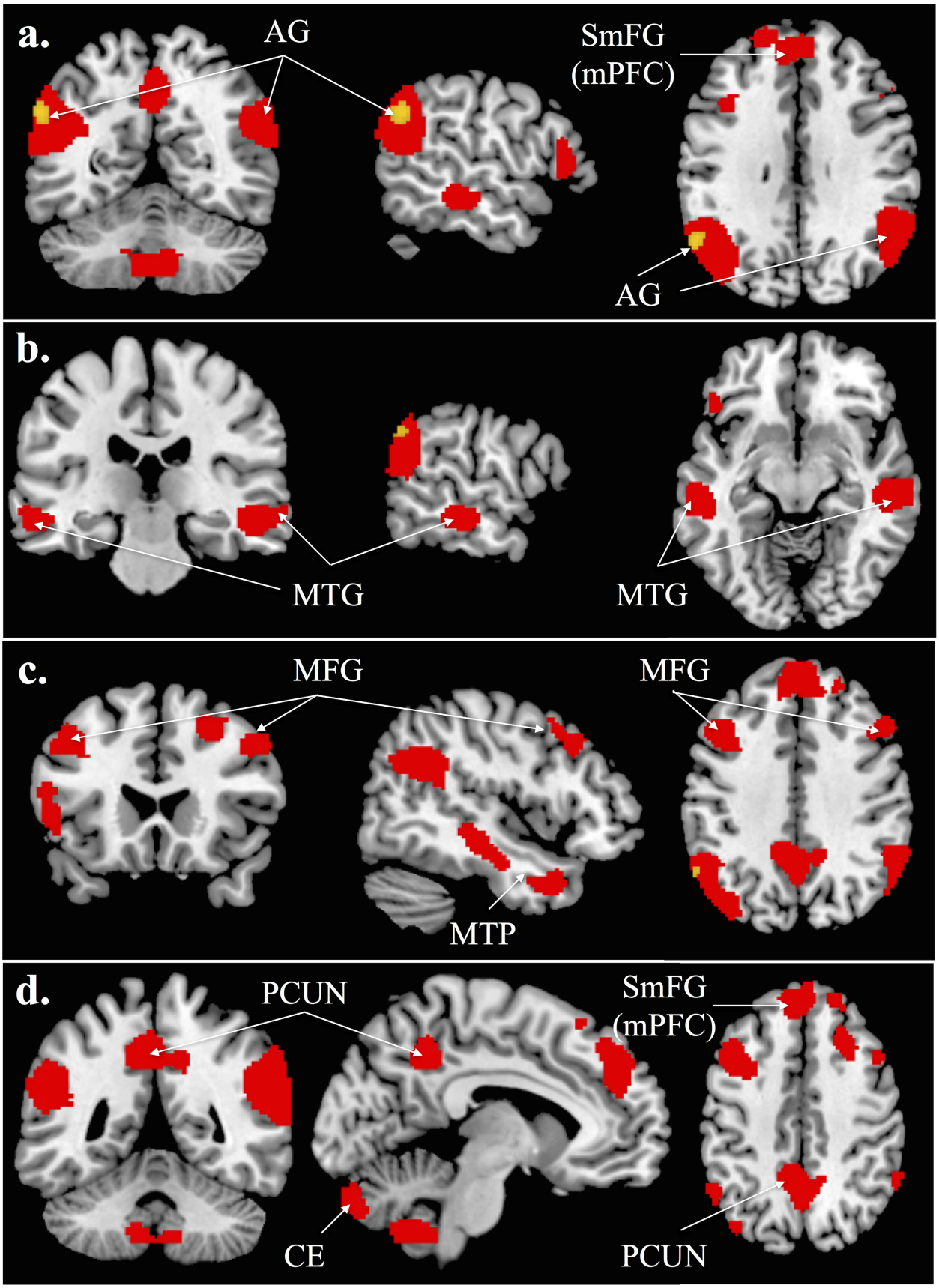

Figure 3: The contrast im-TOM vs. im-NONTOM is depicted in red and ex-TOM vs. ex-NONTOM in yellow, superimposed on a high resolution brain template (the Colin brain) of the MRIcron software (p <.001, cluster extend threshold of 48 voxels, Monte Carlo corrected). Abbreviations: AG: angular gyrus, SmFG: superior medial frontal gyrus, mPFC: medial prefrontal cortex, MTG: middle temporal gyrus, MFG: middle frontal gyrus, CE: cerebellum, PCUN: precuneus. Coordinates: a) MNI: -56, -58, 32, b) MNI: -60, -24, -10, c) MNI: 46, 22, 38, d) MNI: -8, -48, 44.

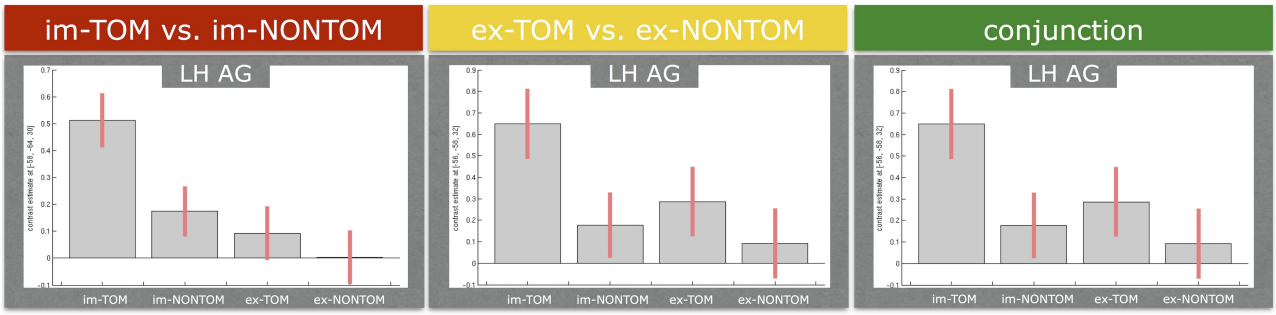

Figure 4: Bar plots of BOLD signal changes (contrast estimates) at the peak voxel of the left (LH) angular gyrus (AG) for *implicit* and *explicit* false belief processing as well as their conjunction. Threshold for the *implicit* and *explicit* contrasts: p <.001 and cluster extend threshold of 48 voxels (Monte Carlo corrected), for the conjunction p <.05 and cluster extend threshold of 120 voxels (Monte Carlo corrected). The error bars represent 90% confidence intervals. Please note that the peak voxel of the LH AG is the same for the explicit contrast and the conjunction, therefore the barplots are also the same.

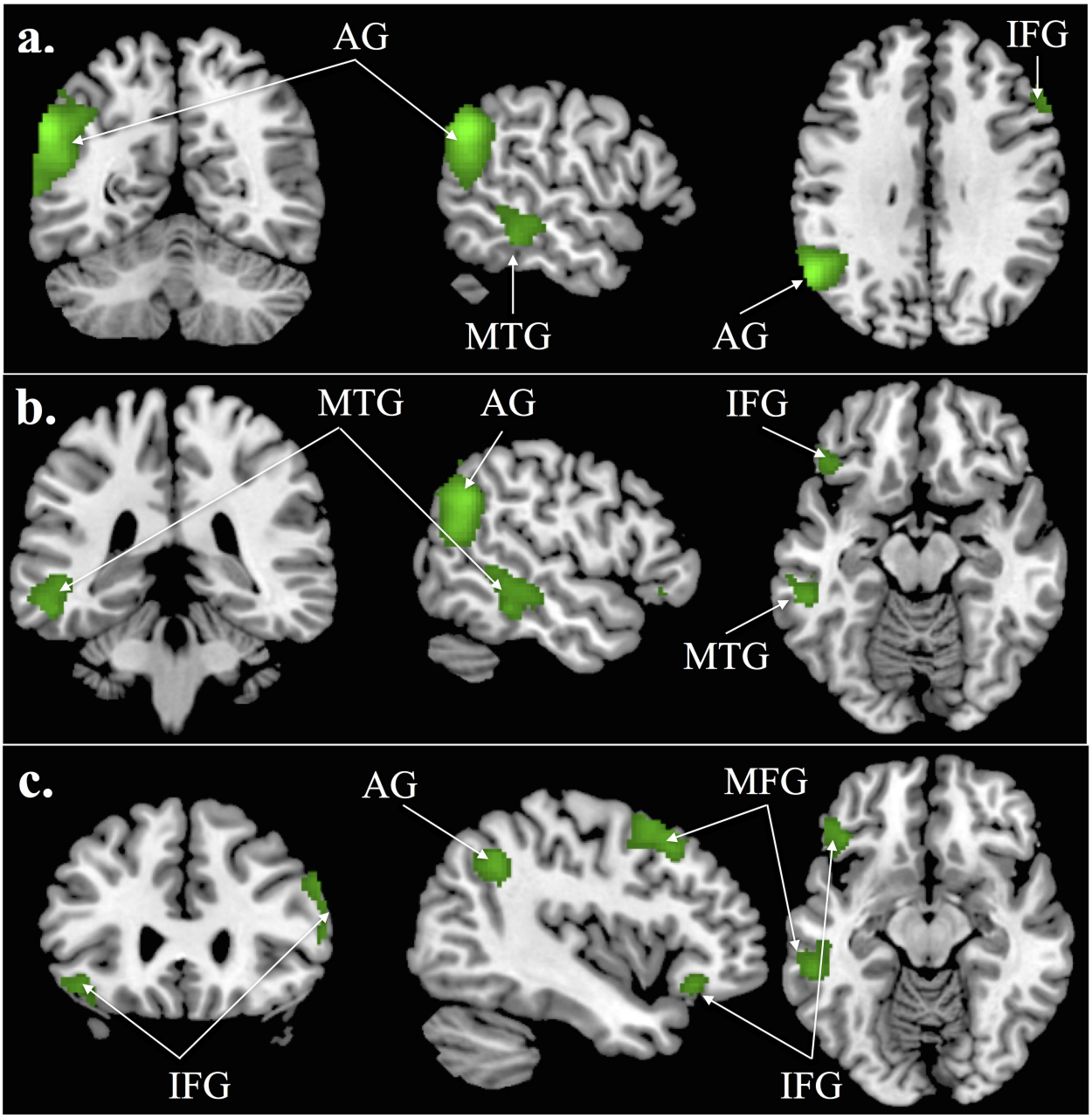

Figure 5: The activations of the conjunction of the contrasts **im-TOM** vs. **im-NONTOM** and **ex-TOM** vs. **ex-NONTOM** are shown in green (conjunction voxel threshold of p <.05 and cluster extend threshold of 120 voxels, Monte Carlo corrected). Abbreviations: AG: angular gyrus, MTG: middle temporal gyrus, MFG: middle frontal gyrus, IFG: inferior frontal gyrus. Coordinates: a) MNI: -56, -58, 32, b) MNI: -52, -46, -14, c) MNI: -42, 28, -12.

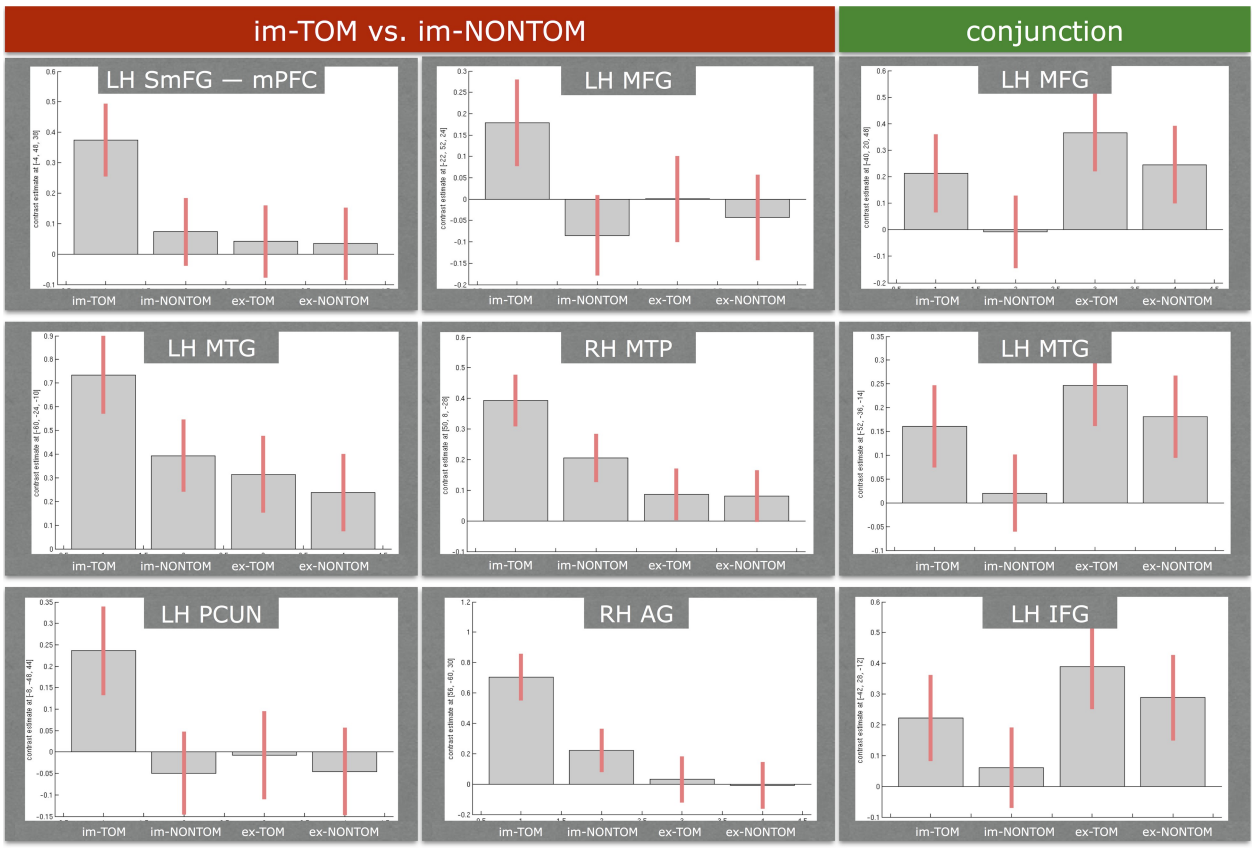

Figure 6: Bar plots of BOLD signal changes (contrast estimates) at the peak voxel of each cluster for *implicit* false belief processing and the conjunction of *implicit* and *explicit* contrasts. Abbreviations: RH: right hemisphere, LH: left hemisphere, AG: angular gyrus, SmFG: superior medial frontal gyrus, mPFC: medial prefrontal cortex, MTG: middle temporal gyrus, MTP: middle temporal pole, MFG: middle frontal gyrus, PCUN: precuneus. Please note that we have placed the plots for precuneus, AG and IFG next to each other for illustration reasons and with absolutely no intention of denoting that they are overlapping regions. For *explicit* and *implicit* contrasts: individual voxel threshold of p <.001 and cluster extend threshold of 48 voxels (Monte Carlo corrected). For the conjunction: individual voxel threshold of p <.05 and cluster extend threshold of 120 voxels (Monte Carlo corrected). The error bars represent 90% confidence intervals.