Do Implicit and Explicit Belief Processing Share Neural Substrates?

Claire K. Naughtin,^{1,*} Kristina Horne,¹ Dana Schneider,² Dustin Venini,¹
Ashley York,¹ and Paul E. Dux^{1,*}

¹*School of Psychology, The University of Queensland, Queensland, Australia*

²*Institute of Psychology, Friedrich Schiller University Jena, Jena, Germany*

Abstract: Humans rely on their ability to infer another person's mental state to understand and predict others' behavior ("theory of mind," ToM). Multiple lines of research suggest that not only are humans able to consciously process another person's belief state, but also are able to do so implicitly. Here we explored how general implicit belief states are represented in the brain, compared to those substrates involved in explicit ToM processes. Previous work on this topic has yielded conflicting results, and thus, the extent to which the implicit and explicit ToM systems draw on common neural bases is unclear. Participants were presented with "Sally-Anne" type movies in which a protagonist was falsely led to believe a ball was in one location, only for a puppet to later move it to another location in their absence (false-belief condition). In other movies, the protagonist had their back turned the entire time the puppet moved the ball between the two locations, meaning that they had no opportunity to develop any pre-existing beliefs about the scenario (no-belief condition). Using a group of independently localized explicit ToM brain regions, we found greater activity for false-belief trials, relative to no-belief trials, in the right temporoparietal junction, right superior temporal sulcus, precuneus, and left middle prefrontal gyrus. These findings extend upon previous work on the neural bases of implicit ToM by showing substantial overlap between this system and the explicit ToM system, suggesting that both abilities might recruit a common set of mentalizing processes/functional brain regions. *Hum Brain Mapp* 38:4760–4772, 2017. © 2017 Wiley Periodicals, Inc.

Key words: implicit theory of mind; fMRI; temporoparietal junction; mentalizing; social cognition

INTRODUCTION

In any given social interaction, humans will likely engage in some degree of mental state processing to infer what their social partner is thinking, feeling or intending to do. This capacity to process another person's internal state is referred to as having a "theory of mind" (ToM), and is crucial for understanding and predicting another person's behavior [Frith and Frith, 2005; Premack and Woodruff, 1978]. ToM is considered to be a key milestone in social development [Razza and Blair, 2009; Slaughter et al., 2013] that persists into adulthood, but can decline later in life [Henry et al., 2013; cf. Maylor et al., 2002]. Impairments in ToM have been linked to numerous developmental disorders, such as autism and schizophrenia, whereby such individuals fail to pass formal tests of this

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: Australian Research Council Special Research Initiative grant for Science of Learning Research Centre; Contract grant number: SR120300015; Contract grant sponsor: Young Researcher Support from the Friedrich Schiller University Jena, Germany; Contract grant number: DRM/2014-02

*Correspondence to: Dr Claire K. Naughtin. E-mail: claire.naughtin@gmail.com or Paul E. Dux. E-mail: paul.e.dux@gmail.com

Received for publication 20 February 2017; Revised 13 June 2017; Accepted 14 June 2017.

DOI: 10.1002/hbm.23700

Published online 23 June 2017 in Wiley Online Library (wileyonlinelibrary.com).

ability and show altered mental state processing [Baron-Cohen et al., 1985; Brune, 2005; Moran et al., 2011]. Although the capacity for ToM was traditionally thought to emerge around 4 years of age [Perner and Lang, 1999; Wellman et al., 2001], more recent evidence suggests that children can spontaneously infer and track another person's mental state at much younger ages [e.g., Scott and Baillargeon, 2017; Senju et al., 2011; Southgate et al., 2007]. Such findings have been taken as evidence for an *automatic* [Cohen and German, 2009] or *implicit* [Apperly and Butterfill, 2009] belief processing system that is distinct from its more explicit counterpart [for recent reviews, see Schneider et al., 2015, 2017]. Here we shed new light on how implicit and explicit ToM processes are related and reflected in the human brain.

Clements and Perner [1994] provided initial evidence for the distinction between implicit and explicit ToM using an anticipatory-looking paradigm. This approach was based on the classic "Sally-Anne" measure of explicit ToM [Wimmer and Perner, 1983], in which a protagonist (Sally) puts a ball in one of two locations and then leaves the scene. While Sally is away, a second character (Anne) moves the ball to the other available location. Sally then returns to the scene, and the critical question is, where will Sally look for the ball? In addition to this explicit measure of ToM, Clements and Perner also assessed children's implicit belief processing by measuring anticipatory eye movements to the two possible ball locations. Children aged 2 years 11 months and older were more likely to look at the location consistent with the protagonist's belief about the ball location, rather than the location where the ball actually was. In contrast, only children aged 4 years and above were able to pass the explicit belief test. Similar tests of implicit belief processing suggest that it is present in children as young as 15 months old [Onishi and Baillargeon, 2005; Senju et al., 2011; Southgate et al., 2007; Surian et al., 2007], although the status of infant implicit ToM is disputed [Ruffman, 2014]. Implicit ToM is also evident in adults [Schneider et al., 2012a,b, 2014a], but seems to be impaired in autism spectrum disorder populations [Schneider et al., 2013; Senju et al., 2009]. These behavioral findings in infants and adults suggest there are fundamental differences between these two forms of belief processing.

Apperly and Butterfill [2009] and Butterfill and Apperly [2013] put forward a two-path ToM system account to explain the developmental dissociation between implicit and explicit forms of mentalizing: An earlier developing pathway that can rapidly extract belief-like states without awareness, and a later developing pathway that supports conscious, deliberate belief processing, and is more susceptible to interference from other executive functions. The former pathway arguably underpins behavior in implicit eye-tracking measures of belief processing and the latter pathway is reflected by explicit mental state judgements. Furthermore, it has been hypothesized that the implicit

ToM system represents the minimal elements of a belief that can be used to track another person's internal state, without the need to generate a complete belief representation [Butterfill and Apperly, 2013], or forms part of our core "social sense" [Kovács et al., 2010]. In support of this idea, Schneider et al. [2012a] found individuals' ability to implicitly monitor a protagonist's belief about a ball location sustained over a 1-h testing session. This suggests that, unlike other automatic cognitive processes (e.g., stimulus-driven attentional capture; Asplund et al. [2010]), implicit belief monitoring does not rapidly habituate over time. However, Low and Watts [2013] found the implicit ToM system is limited in the types of belief states it can process, such that children aged 3–4 years and adults alike can monitor another's belief about an object's location, but not beliefs about an object's identity.

A host of studies have used functional magnetic resonance imaging (fMRI) to explore the neural basis of the explicit ToM system [for recent meta-analyses, see Moltenberghs et al., 2016; Schurz et al., 2014]. One way to probe explicit ToM processes is to compare brain activity under conditions where participants are asked to explicitly evaluate the mental state of a protagonist, compared with making a physical judgement. Such comparisons reliably produce increased activity in the medial prefrontal cortex (mPFC), temporoparietal junction (TPJ), superior temporal sulcus (STS), precuneus, and temporal poles [e.g., Dodell-Feder et al., 2011; Gallagher et al., 2000; Saxe and Kanwisher, 2003]. Moreover, these brain regions can also distinguish between belief states that are consistent with the state of affairs (i.e., a true-belief), relative to those that are inconsistent (i.e., a false-belief), suggesting that they are sensitive to the content rather than merely the presence of belief states [Aichhorn et al., 2009; Döhl et al., 2012; Sommer et al., 2007]. In particular, the TPJ is considered to be a core component of the explicit mentalizing network (specifically in the right hemisphere), and is involved in a range of mental state reasoning tasks [Saxe, 2009; Saxe and Kanwisher, 2003]. Indeed, a recent fMRI meta-analysis of ToM studies found the TPJ is consistently engaged in ToM processing regardless of the specific task demands [Schurz et al., 2014].

Comparatively, much less is known about the neural bases of implicit ToM processes and the extent to which this system overlaps with the explicit ToM system. Imaging work to date into implicit ToM has relied on indirect measures or paradigms that were designed to elicit spontaneous mental state processing [e.g., Castelli et al., 2000; Gobbi et al., 2007; Iacoboni et al., 2004; Ramsey et al., 2013; Wagner et al., 2011; Wheatley et al., 2007]. As a recent example, Rice and Redcay [2016] engaged participants in an interaction with another social partner, which they were led to believe was shown either in real time or prerecorded. These authors found the bilateral TPJ, precuneus, dorsomedial prefrontal cortex (dmPFC), mPFC, and ventromedial prefrontal cortex (vmPFC) activated more

strongly when participants thought the interaction was in real-time rather than pre-recorded. Hence, live social interactions could place greater demands on the (implicit) mentalizing system or engage a broader selection of social cognitive processes (e.g., speech processing) to create a shared social experience [Rice and Redcay, 2016]. Also, works employing violation of expectation paradigms [e.g., Bardi et al., 2017; Kovács et al., 2014] have suggested that the TPJ is specifically engaged in processing implicit mental states. Note, however, these findings have been called into question given that the belief state processing was confounded with an attention-checking procedure in these paradigms [Phillips et al., 2015].

While the brain regions implicated in the aforementioned studies somewhat overlap with those implicated in explicit ToM processing, the extent to which this activity reflects implicit belief processing is unclear for two reasons. First, participants in these experiments passively viewed the stimuli used to evoke mental state processing or were instructed to make judgements about its content. Put differently, the belief-inducing stimuli were always task-relevant. Even though participants were never explicitly asked to report mental states in these studies, there is no way to guarantee that they did not still engage in explicit, conscious ToM processing. Second, as participants' level of engagement in mental state processing was never directly assessed, we do not know whether they were unaware of the mentalizing nature of the experimental stimuli. Therefore, the apparent overlap could simply reflect a single explicit ToM process, rather than a common network for implicit and explicit ToM abilities. We have previously developed ways to address these two concerns [Schneider et al., 2014b]. First, we made the belief-evoking stimuli task-irrelevant by engaging participants in a secondary tone detection task, and second, we explicitly probed participants' level of belief processing using a funneled debriefing questionnaire [Bargh and Chartrand, 2009] to identify (and exclude) participants who were aware of the belief manipulations.

Specifically in our work [Schneider et al., 2014a], we used Sally-Anne type movies that were adapted from earlier developmental studies [Onishi and Baillargeon, 2005; Senju et al., 2011; Southgate et al., 2007] to be appropriate for an adult sample [Schneider et al., 2012a]. In the initial belief set-up phase, participants observed a protagonist watch a puppet move a ball to one of two possible boxes, before leaving the room, either before the puppet made a second ball movement (true-belief condition) or after (false-belief condition). Then in the belief test phase, while the actor was out of the room, the puppet moved the ball to the opposite box location for a third time, meaning the ball was in the original box location when the actor then returned. The actual ball location therefore either matched or mismatched the protagonist's belief about its location (i.e., they had a true- or false-belief, respectively). Critically, activity was compared between true- and false-belief

conditions during the belief test phase where the videos were physically identical. Thus, unlike in previous paradigms, any differences in activity could not be driven by visual differences [e.g., Gobbi et al., 2007; Iacoboni et al., 2004; Wheatley et al., 2007]. We found the precuneus and left STS were more responsive to false- than true-belief scenarios, suggesting that these brain areas constitute the core components of the implicit ToM processing system.

A notable contrast between the findings from Schneider et al. [2014a] and previous fMRI studies on implicit ToM processing is the lack of TPJ involvement. This result is surprising given that this brain region appears to be a core component of the explicit ToM network [Saxe, 2009; Saxe and Kanwisher, 2003; Schurz et al., 2014]. Specifically, Schneider et al. [2014a] found activity in the TPJ could distinguish between the presence and absence of explicit belief states, but was not modulated by true and false implicit belief conditions. This finding suggests that this region might not be sensitive to the differential demands placed on the implicit ToM system and/or the content of implicit belief states.

Hyde et al. [2015] aimed to resolve this discrepancy using functional near-infrared spectroscopy, a neuroimaging technique that has greater temporal resolution (albeit reduced spatial resolution) than fMRI and can identify more fine-grain changes in brain activity across time. These authors modified the paradigm used by Schneider et al. [2014a] to also include trials where the protagonist could directly perceive the ball (i.e., the two possible box locations were transparent). The direct perception condition was comparable to the true-belief condition in terms of the protagonist's mental state (their belief about the ball location matched reality), and to the false-belief condition in terms of the physical movements (i.e., to control for low-level visual differences between conditions, such as in Schneider et al. [2014a]). Hyde et al. [2015] found greater activity in the right TPJ for false-belief trials, relative to both true-belief and direct perception trials, specifically around the time period where the puppet moved the ball without the protagonist's knowledge (i.e., when the false-belief was created). It is unclear from these findings, however, whether Schneider et al. [2014a] failed to find evidence of implicit ToM processing in the TPJ due to limitations in the temporal resolution of fMRI (as suggested by Hyde et al. [2015]), or whether this result reflects a functional difference in this brain region's role in implicit and explicit belief processing.

Thus, based on the work to date, it is still unclear how implicit belief processing is reflected in the brain. All previous fMRI investigations into spontaneous, implicit belief processes have relied solely on comparisons between true- and false-belief conditions [Bardi et al., 2017; Hyde et al., 2015; Kovács et al., 2014; Schneider et al., 2014a], which might have underestimated or overlooked a more general belief processing system in the brain that is recruited for both explicit and implicit ToM tasks. Studies examining

the neural bases of goal, intention and trait attribution, social inference processes—which also draw heavily on ToM operations and perhaps even those implicated in false-belief analysis—are indeed suggestive of a common belief processing system. These studies have found that implicit, explicit, and attributional processes tap overlapping brain regions (e.g., TPJ, mPFC) and have a similar temporal signature as assessed via electroencephalogram (EEG) [Van Overwalle and Vandekerckhove, 2013].

To test whether a common belief processing system specifically exists for ToM operations, the present study employed a similar Sally–Anne paradigm to Schneider et al. [2014a]; see also Hyde et al. [2015] in conjunction with fMRI. On each trial, participants watched movies involving a protagonist, a puppet, and a ball. The protagonist sometimes watched the puppet move the ball between two possible locations, meaning their belief about the ball location was later falsified when the puppet switched the ball location in their absence. Alternatively, they would have their back to the entire scenario and thus, never developed any beliefs about the ball's location or the puppet.

We used a region of interest (ROI) approach to identify brain regions involved in spontaneous, implicit belief processing. To this end, we compared brain activity between conditions where the protagonist held a false-belief about the state of affairs, relative to when they had no preconceived beliefs about the scenario. Our ROIs reflected brain areas that are sensitive to *explicit* belief states, and were independently localized by comparing activity between conditions that require conscious mental state processing to conditions devoid of any belief states [e.g., Dodell-Feder et al., 2011]. If the implicit ToM system draws on the same general-belief brain regions as the explicit ToM system, we should see a large degree of overlap between these two functional networks. In particular, if the TPJ does indeed form part of the functional network of implicit ToM, then we would expect it would differentially respond to false-belief situations, relative to situations where the protagonist has no preconceived beliefs.

METHOD

Overview

Participants were presented with false-belief and no-belief movies that consisted of two phases: First, a belief set-up phase where we manipulated whether the protagonist had a false-belief or no-belief about the location of a ball, and second, a belief test phase where we assessed for behavioral and neural differences between the two belief conditions, whilst keeping all physical elements of the display constant (Fig. 1). The belief test phase of both belief conditions was identical (the exact same movie was used), except for the fact that the protagonist either had a false-belief or no-belief about the ball location based on what

they experienced in the belief set-up phase (where we manipulated the specific movie events between conditions to induce different belief expectations). Therefore, any changes in behavior or brain activity during this belief test phase would be driven by the differential belief processing demands alone. In the false-belief condition, there was a mismatch between their belief and the protagonist's belief about the ball location, meaning that participants had to simultaneously monitor multiple conflicting beliefs (higher mentalizing demands). In the no-belief condition, however, the protagonist did not hold any pre-existing beliefs about the ball location (as they never witnessed the ball being moved), meaning that the participant only had to monitor their own belief state (lower mentalizing demands).

The final frame of the belief test movie froze for 6 s at the end of the phase so that we could monitor participants' anticipatory eye movements as an implicit measure of belief processing. This approach has been used in previous implicit ToM experiments that compared true- and false-belief conditions: the logic being that participants are more likely to first fixate (and to fixate for longer) on the box that does not contain the ball if the protagonist holds a false-belief that the ball is at that location, relative to a true-belief that the ball is not at that location (Schneider et al., 2012a,b, 2013, 2014a,b). This hypothesis is based on the assumption that preferential looking to the protagonist's belief location is indicative of implicit mental state attribution (Onishi and Baillargeon, 2005; Southgate et al., 2007; Surian et al., 2007). We expected to see a similar increase in no-ball fixations under the false-belief compared with the no-belief condition.

Throughout both movie phases, participants' task was to detect the presence of high- and low-frequency tones, meaning that the movies were task-irrelevant and their task did not require active engagement in any belief processing. At the end of the experiment, participants also completed a funneled debriefing questionnaire to assess for awareness of belief processing (adapted from Bargh and Chartrand [2009]). All participants identified as being conscious of our belief processing manipulation were excluded from further data analyses.

Participants

We recruited 33 volunteers (13 males) with a mean age of 24.4 years ($SD = 4.0$) for this experiment. All participants had normal or corrected-to-normal vision, gave written informed consent, had no history of neurological or psychiatric disorders, and were financially reimbursed for their time and participation. Data were excluded from participants who showed evidence of explicit belief processing in the debriefing questionnaire ($n = 10$) and from one participant that had no behavioral data recorded due to a technical error (henceforth, $N = 22$). The University of Queensland's Human Research Ethics committee approved this experimental protocol.

a) Trial structure

b) Belief set-up sequence

c) Belief test sequence

Figure 1.

Schematic representation of the implicit ToM task. (A) Overview of trial structure. Participants are first presented with a “get ready” prompt, followed by the belief set-up movie sequence and the belief test movie sequence, with an extended 10 s inter-stimulus interval between each movie. (B) The belief set-up sequence was used to manipulate the implicit belief state of the protagonist in the movie. On false-belief trials, the protagonist faced forward and watched as the puppet moved the ball to one of the box locations, and then to the other, before leaving the scene. On no-belief trials, the puppet made the same two ball movements, except the protagonist had their back to the puppet. The protagonist therefore never had the opportunity to develop any pre-existing beliefs about the puppet or the ball’s location. (C) The belief test sequence was identical across both

false-belief and no-belief trials, and was designed to test the impact of the presence or absence of the protagonist’s prior belief about the ball’s location. While the protagonist was still out of the room, the puppet moved the ball from its current location to the other box, and then the protagonist returned to sit at the table facing forward. On the false-belief trials, the protagonist’s belief about the ball’s location (e.g., left box) was incongruent with its actual location (e.g., right box), but on no-belief trials, the protagonist still had no pre-existing beliefs about the ball’s location, or any knowledge that the ball exists at all. The belief test sequence ended with the final movie frame frozen for 6 s, during which time we measured anticipatory eye movements to each of the box locations and the protagonist’s face.

Experimental Tasks

Implicit ToM task

Upon entering the MRI scanner, participants first completed the implicit ToM task adapted from our previous fMRI study [Schneider et al., 2014a]. As previously described (see Overview), each trial consisted of a belief set-up movie followed by a belief test movie to assess the effects of the prior belief manipulation (which was the same for each condition). The false-belief set-up phase began with the protagonist facing towards the camera with a ball in the center of the table (Fig. 1B, upper panels; www.youtube.com/watch?v=2tG7SnwKi1E). They watched

as a puppet appeared and placed a central ball in one of the boxes, and then moved the ball to the other box (the left and right movements were randomized across trials). This phase concluded with the protagonist leaving the room. In the no-belief set-up phase, the puppet’s behavior is identical (it made the same two ball movements), but the protagonist had her back to the puppet throughout the entire movie (Fig. 1B, lower panels; https://youtu.be/LI_WJXLepB0). Thus, the protagonist had no opportunity to create any beliefs about the ball, its location, or the puppet. The same movie was used in the belief test phase for false-belief and no-belief trials (Fig. 1C; <https://youtu.be/evFRneJxxYE>), in which the puppet returned, moved the ball from its

current box to the opposite box and then left the scene. The protagonist returned and sat at the table, facing forward with her gaze centered between the two boxes. The belief test phase ended with the final frame frozen for a further 6 s, during which time we analyzed anticipatory eye movements to the protagonist's face and the two box locations.

Participants were instructed to pay attention to the movies, but their primary task was to detect the presence of high- and low-frequency tones (2 and 1.1 kHz, respectively) by pressing one of two buttons as quickly and accurately as possible. This task was designed to distract participants from the content of the movies, thus minimizing the extent to which they explicitly processed the belief states of the protagonist or puppet. There were four tones presented in the belief set-up phase and two tones in the belief test phase. The timing of these presentations was randomized across both phases, with the constraints that there was a minimum of 4 s between each tone and that no tones appeared during the 6 s still frame of the belief test movie (where we analyzed eye movements). Performance on this task was close to ceiling, where participants correctly identified the tones on an average of 94.2% of trials ($SD = 6.7\%$). Each trial lasted for 74 s and consisted of a "get ready" prompt (2 s), followed by the belief set-up movie (26 s), an interstimulus interval (10 s), the belief test movie (26 s), and a second interstimulus interval (10 s) (Fig. 1A). The extended delay between each stimulus ensured that the BOLD signal returned to baseline before the next stimulus appeared. There were a total of 20 false-belief and 20 no-belief trials split over 8 runs. The protagonist wore a visor throughout all the movies to ensure that her eye gaze did not produce any attentional cueing effects [Bayliss and Tipper, 2006], and we have previously confirmed that this is not a contributing factor in our spontaneous, implicit belief paradigm [Schneider et al., 2012a].

Funneled debriefing questionnaire

Immediately following the implicit ToM task, participants completed a funneled debriefing questionnaire to assess their level of explicit belief processing in the proceeding task. This 10-item questionnaire was similar to what we have used previously [Schneider et al., 2012a, 2014a], and is designed to incrementally probe participants' level of awareness of our belief manipulation (e.g., "what did you think the story was in the videos?"). We excluded any participants whose responses suggested that they had explicit awareness of our belief conditions and/or engaged in belief processing during the task. That is, if they specifically referred to the mental state or the expected behavior of the protagonist or puppet (e.g., "the koala [puppet] was trying to trick the girl [protagonist]"). Importantly, this questionnaire was completed after the implicit measure of ToM was acquired, meaning that any mention of explicit mental states could not interfere with participants' level of awareness during that task. As

mentioned in the Participants section, 10 participants did not pass this assessment, as their responses reflected awareness of our belief manipulation and/or engagement in explicit mental state processing, and were excluded from further data analyses.

Explicit general-belief localizer task

The final part of the fMRI scanning session consisted of an explicit general-belief localizer task adapted from Dodell-Feder et al. [2011]. In this task, participants were presented with short vignettes, followed by a statement that they identified as "true" or "false" by pressing one of two buttons on the button box. The vignettes either described a scenario that required participants to think about the updated mental state of the character(s) (i.e., a false-belief trial) or the physical or logical aspects of an updated scenario (i.e., a false-photograph trial). For example, a false-belief vignette could be "expecting the game to be postponed because of the rain, the Garcia family took the subway home. The score was tied, 3–3. During their commute the rain stopped and the game soon ended with a score of 5–3." The associated statement was, "The Garcia family arrives home believing the score is 5–3." On the other hand, an example false-photograph vignette would be "accounts of the country's bustling economic success were recorded in both fiction and non-fiction books from the early 1900s. Soon after, a horrible plague hit the country and the country was sent into an economic depression." The associated statement was, "Early 1900s novels portray the country as experiencing economic wealth." The correct response to these two statements would be "false" and "true," respectively. Each trial lasted for 32 s, and consisted of the presentation of the vignette (14 s), followed by the response statement (4 s), and an intertrial interval (14 s). There were a total of 20 false-belief trials and 20 false-photograph trials split over 2 runs.

Auxiliary measures

At the end of the scanning session, participants completed three self-report questionnaires to identify those who present with social processing difficulties, which could influence their performance on either ToM task. These measures included the Autism-Spectrum Quotient (AQ) [Baron-Cohen et al., 2001] to assess for traits associated with the autistic spectrum, the Empathising Quotient (EQ) [Baron-Cohen and Wheelwright, 2004] to assess abilities associated with understanding, predicting, and experiencing another person's emotions, and the Systemizing Quotient (SQ) [Baron-Cohen et al., 2003] to assess abilities associated with understanding, analysing and predicting external elements and systems. All participants fell within the normal range on the AQ ($M = 19.2$, $SD = 6.1$, where the clinical cut-off is 32 out of 50), EQ ($M = 47.6$, $SD = 10.8$ out of 80) and SQ ($M = 69.7$, $SD = 21.9$ out of 80), and thus, we will not discuss these measures any further.

fMRI Data Acquisition

We acquired anatomical and functional images using a 3 T Siemens Trio MRI scanner (Erlangen, Germany) with a 32-channel head coil. Participants wore MRI-compatible headphones, lay supine in the scanner and viewed the visual display via a rear-projection mirror mounted on the head coil. We monitored participants' eye movements during the implicit ToM task using an Eyelink 1000 (500 Hz) eye-tracker mounted at the bore of the scanner. All functional T2*-weighted images were aligned to the anterior-posterior commissure and measured using the following GRE EPI sequence parameters: TR = 2 s, TE = 25 ms, FA = 90°, FOV = 192 × 192, matrix = 64 × 64, in-plane resolution = 3 mm). Each functional volume consisted of 33 slices with a 3 mm thickness (10% inter-slice gap), and its acquisition was synchronized with the stimulus presentation. A total of 189 volumes and 164 volumes were measured for each run of the implicit ToM and explicit general-belief localizer tasks, respectively, where the first 3 volumes of each run were discarded from data analyses to allow for T1 equilibrium. A high-resolution T1-weighted image was acquired in the middle of the scanning session using the following MPRAGE sequence parameters: TR = 1.9 s, TE = 2.3 ms, FA = 9°, FOV = 192 × 230 × 256, resolution = 1 mm³.

fMRI Data Analysis

We preprocessed and analyzed the fMRI data using BrainVoyager QX 2.6 (Brain Innovation) and custom MATLAB code. The data preprocessing steps included 3D motion correction (where each functional run image was aligned to the first run), slice-scan-time correction, high-pass temporal filtering (3 cycles/run), and Talairach space transformation [Talairach and Tournoux, 1988]. We analyzed data from the implicit ToM task using an ROI approach, and these ROIs were localized using data from the explicit general-belief localizer task. To isolate these ROIs, we conducted a whole-brain random-effects general linear model (GLM) with regressors defined for the false-belief and false-photograph conditions. The events in each regressor were defined as the nine volumes corresponding to the presentation of the vignette and question and convolved with a double-gamma hemodynamic response function. This analysis approach followed the recommendations of Dodell-Feder et al. [2011].

We contrasted activity between false-belief and false-photograph conditions in the localiser group GLM to isolate the following brain regions at a statistical threshold of $q(\text{FDR}) < 0.05$ (Fig. 2A; peak Talairach XYZ coordinates in parentheses): middle frontal gyrus (MFG; left, -29, 16, 37; right, 23, 22, 41), precuneus (-1, -56, 34), TPJ (left, -46, -60, 21; right, 49, -54, 21), STS (left, -56, -22, -7; right, 53, -23, -4), temporal pole (left, -55, 5, -18; right, 52, 6, -16), dmPFC (-2, 53, 18), and vmPFC (0, 49, -11). These ROIs consist of brain regions that typically activate during explicit ToM tasks [e.g., Aichhorn et al., 2009; Dodell-

Feder et al., 2011; Saxe and Kanwisher, 2003; Van Overwalle, 2009], and were defined based on previously published Talairach coordinates from Schneider et al. [2014a] and Schurz et al. [2014]. Using these ROIs, we extracted individual-subject BOLD time courses from the belief test phase of the implicit ToM task (where stimuli were physically identical across both belief conditions) for false-belief and no-belief trials. Percentage signal change was calculated relative to one volume prior to the onset of the belief test movie. We defined peak amplitude as the greatest percentage signal change during the time window corresponding to 6–20 s following the movie onset [Schneider et al., 2014a].

RESULTS

Behavioral Results

To analyze eye movements from the implicit ToM task, we measured the location and duration of fixations corresponding to five key areas of interest: The protagonist's face, left arm, right arm, and the left and right boxes [Schneider et al., 2012a]. Eye movements were collapsed across the arm and box regions to form the region corresponding to the ball location (the box where the ball was actually located) and the no-ball location (the empty box). Data were analyzed over the time period corresponding to the 6 s still frame of the belief test phase, and we measured the percentage of first fixations and fixation durations at each of the three visual regions (face, ball, and no ball locations). We have previously found that false-belief scenarios enhanced looking behavior at the no-ball location, compared to true-belief scenarios [e.g., Schneider et al., 2012a,b]. To reiterate, we expected participants to look at the no-ball location more under false- versus no-belief conditions.

Counter to our hypothesis, a 2 (Belief Condition: false-belief, no-belief) × 3 (Visual Region: face, ball or no-ball location) repeated-measures ANOVA of the percentage of first fixations revealed a main effect of Visual Region, $F(1, 23) = 28.04$, $\text{MSE} = 2313$, $P < 0.001$, $\eta_p^2 = 0.57$, but no main effect of Belief Condition or interaction, $F_s < 1$ (Table 1). Follow-up t tests revealed that participants were more likely to fixate first on the ball location, relative to both the face and no-ball regions, $t_s > 3.54$, $P_s < 0.002$, which were also significantly different from one another, $t(21) = 3.65$, $P = 0.001$. A similar pattern of results was observed for the analysis of percentage of fixation durations: Main effect of Visual Region $F(1, 24) = 69.0$, $\text{MSE} = 962$, $P < 0.001$, $\eta_p^2 = 0.77$, but no other significant effects or interactions, $F_s < 1$. This main effect was driven by participants fixating longer on the face region, relative to both the ball and no-ball locations, $t_s > 8.48$, $P_s < 0.001$, which did not differ from each other, $t(21) = 1.41$, $P = 0.175$. The relative amount of time participants spent looking at the no-ball location was comparable to our previous studies using this

Figure 2.

Imaging results from regions of interest (ROI) analysis. (A) Explicit theory of mind (ToM) regions localized from the explicit general-belief localizer task. The anatomical image shows the random-effects general linear model for all subjects at a statistical threshold of $q(\text{FDR}) = 0.05$. ROIs were isolated by contrasting activity between false-belief and false-photograph conditions (Dodell-Feder et al., 2011), and were defined by Talairach coordinates taken from relevant prior studies (Schneider et al., 2014; Schurz et al., 2014). The explicit ToM network included the

(1–2) left and right middle frontal gyrus (MFG), (3) precuneus, (4–5) left and right temporoparietal junction (TPJ), (6) dorsomedial prefrontal cortex (dmPFC), (7–8) left and right superior temporal sulcus (STS), (9–10) left and right temporal pole (TempP), and (11) ventromedial prefrontal cortex (vmPFC). (B) Mean peak signal amplitude as a function of Belief Condition (false-belief, no-belief) and ROI. Error bars denote standard error of the mean. * $P < 0.05$, follow-up t tests from significant Belief Condition $\times$ ROI ANOVA.

paradigm [e.g., Schneider et al., 2012a,b], suggesting that the introduction of the no-belief condition did not disrupt participants' typical looking behavior during this task. Surprisingly, using a standard paradigm for eliciting implicit belief processing, we failed to find a behavioral difference in eye movement behavior between false- and no-belief conditions. Possible reasons for this null result are laid-out in Discussion.

fMRI Results

We compared peak BOLD activity from the belief test phase of the implicit ToM task between false-belief and no-belief conditions across each explicit ToM ROI. Evidence for a region being involved in mental state processing would be indexed by greater peak amplitude for false-belief scenarios (where the protagonist's belief about the location of a ball was incongruent with the actual ball's

location), relative to no-belief scenarios (where the protagonist had no pre-existing beliefs about the ball or puppet). A 2 (Belief Condition: false-belief, no-belief) $\times$ 11 (ROI: left MFG, right MFG, precuneus, left TPJ, right TPJ, left STS, right STS, left temporal pole, right temporal pole, dmPFC, vmPFC) repeated-measures ANOVA of BOLD peak amplitude revealed a main effect of ROI, $F(5, 115) = 11.94$, $\text{MSE} = 0$, $P < 0.001$, $\eta_p^2 = 0.36$, and a significant Belief Condition by ROI interaction, $F(4, 84) = 3.28$, $\text{MSE} = 0$, $P = 0.015$, $\eta_p^2 = 0.14$. The main effect of Belief Condition did not reach significance, $F(1, 21) = 1.94$, $\text{MSE} = 0$, $P = 0.178$, $\eta_p^2 = 0.09$. Follow-up t tests revealed significantly greater activity for false-belief trials, compared with no-belief trials, in left MFG, precuneus, right TPJ, and right STS, $t_s > 2.22$, $P_s < 0.038$ (Fig. 2). These differences were consistently observed across all subjects in our sample (Supporting Information, Fig. 1 for distributional information). All other ROIs showed no significant

TABLE I. Behavioral results from implicit ToM task

	First fixation (%)			Fixation duration (%)		
	Face	Ball	No ball	Face	Ball	No ball
False belief	26.7 (24.9)	63.3 (27.6)	9.0 (7.0)	67.0 (20.4)	17.9 (10.6)	15.1 (12.5)
No belief	27.4 (24.0)	64.9 (25.0)	7.4 (7.8)	67.3 (18.6)	17.5 (11.1)	15.2 (10.0)

Mean percentage of first fixations and fixation durations as a function of Belief Condition (false-belief and no-belief) and Visual Region (face, ball, and no-ball locations). Standard deviations are listed in the parentheses.

difference, $t_s < 1.78$, $P_s > 0.090$. Extending on the small group of brain regions identified in initial investigations into spontaneous, belief-specific implicit ToM abilities [Bardi et al., 2017; Hyde et al., 2015; Kovács et al., 2014; Schneider et al., 2014a], these results demonstrated substantial overlap between the implicit and explicit ToM neural systems and point toward a general system in the brain for belief processing.

DISCUSSION

This study had two key objectives. First, to explore whether there is a general belief processing system in the brain that is recruited by the implicit and explicit ToM systems. Second, to resolve the current conflict in the literature surrounding the role of the TPJ in implicit ToM processes. Using tightly controlled scenarios that encouraged processing of a protagonist's false-belief about a ball's location, without participants' awareness, we found evidence of implicit belief processing in a large number of brain regions isolated via a classic explicit ToM paradigm. These brain regions were more responsive under false-belief scenarios, relative to no-belief scenarios, and included the right TPJ, right STS, precuneus, and left MFG. These differences were observed despite no physical differences between the stimuli used in each belief condition; the only factor that distinguished the two scenarios was the presence or absence of a false-belief state in the initial belief set-up phase. These findings expand upon the previous fMRI investigations conducted into the implicit ToM system [Hyde et al., 2015; Schneider et al., 2014a], and demonstrate a common subset of brain regions that support both forms of mental state processing. In addition, the present results converge with previous work that has found common overlap between the neural bases and temporal dynamics of implicit and explicit processing of goals, intentions, trait attributions, and social inferences [van Overwalle and Vandekerckhove, 2013]. However, there are still some regions that appear to be specifically recruited for explicit ToM processes (or at least those mentalizing processes that are only tapped in verbal scenarios as opposed to visual animations as used in our iToM conditions), as not all explicit ToM brain regions were modulated by the presence of implicit false beliefs. This finding further supports the proposed theoretical and developmental distinction between explicit and implicit systems

that underlie domain-general (stimulus modality invariant) ToM operations [Apperly and Butterfill, 2009; Clements and Perner, 1994].

Despite the clear distinction between false- and no-belief conditions in our imaging results, we failed to find any behavioral difference between our belief conditions. Specifically, participants spent the same amount of time looking at each of the box locations, regardless of whether the protagonist had a false-belief about the ball's location, or no belief at all. Critically, this was not due to a systematic change in amount of time participants spent looking at the no-ball location under the no-belief condition, as participants spent a similar amount of time looking at this location to what we typically see in this paradigm [e.g., Schneider et al., 2012a,b]. It should be noted that the majority of studies into implicit ToM have compared behavioral measures between true- and false-belief conditions [Onishi and Baillargeon, 2005; Schneider et al., 2012a; Senju et al., 2011; Surian et al., 2007], or have had no behavioral measure at all [Hyde et al., 2015]. Our findings, therefore, reflect the first behavioral measure of false-belief processing relative to a no-belief baseline.

One possible explanation for the dissociation of our behavioral and imaging results could be sensitivity of each measure (i.e., fMRI might be better able to detect changes driven by an actor's belief state than eye movements). Alternatively, these findings could suggest that true-belief scenarios, rather than false-belief scenarios, were the key condition that drove the previously observed change in looking behavior: Participants spent *less* time, relative to false-belief/baseline conditions, looking at locations that were not socially relevant when the social partner's belief matched the actual state of affairs. Indeed, individuals are more likely to return their gaze away from a peripheral object back to a protagonist in joint attention scenarios when the protagonist consistently reciprocates the participants' social bids, relative to when they consistently reject them [Bayliss et al., 2012]. These findings suggest that peripheral objects might be inherently curious and attention grabbing, but they can lose their significance if another person's behavior, or belief, deems them to be less socially relevant.

Evidence for the role of the TPJ in the implicit ToM system is mixed. Indeed, some studies have found that the right TPJ was more active when a protagonist's belief was

a false, relative to a true, reflection of reality [Bardi et al., 2017; Hyde et al., 2015; Kovács et al., 2014], whereas others have failed to find any difference despite using multiple analytic approaches [Schneider et al., 2014a]. The results from this study suggest that the right TPJ is indeed part of the implicit ToM system, as it can detect the presence versus absence of a false belief state. Similar conflicting findings have been noted in the explicit ToM literature, whereby the right TPJ consistently shows greater activity for false-belief scenarios versus false-photographs or no-belief control trials [Dodell-Feder et al., 2011; Saxe and Kanwisher, 2003], but does not reliably distinguish between true- and false-belief conditions [Aichhorn et al., 2009; Döhnell et al., 2012; Sommer et al., 2007]. Our TPJ ROI corresponds to the more dorsal/posterior part of the TPJ, which typically activates more strongly during ToM tasks that contrast false-beliefs and false-photographs (precisely the task we used for our explicit ToM localizer [Dodell-Feder et al., 2011]), contrary to the more ventral/anterior part of the TPJ, which was more responsive during ToM tasks that involve social animations, mind-in-the-eyes judgements or rational reasoning [Schurz et al., 2014]. A recent meta-analysis supported this functional distinction between the TPJ regions in showing that the more anterior region of the TPJ is involved in both attentional and ToM processes, whereas the more posterior region is specific to the social domain [Krall et al., 2015]. The TPJ ROI used by Kovács et al. [2014] and Hyde et al. [2015] was slightly more anterior (e.g., XYZ = 47, -49, 18, Talairach coordinates) to the ROI used in our current and previous work [Schneider et al., 2014a]. Differences in the functional localization of the TPJ might account for these discrepant TPJ results, and future research could usefully explore whether the anterior and posterior sections of the TPJ are differentially involved in implicit belief processing.

Differences in participants' task are another factor that could account for these discrepant TPJ results. In our work, participants' primary task is to detect high and low tones, with the aim of distracting them from (or at least reducing the chances of) engaging in explicit mentalizing while watching the videos. Conversely, Kovács et al. [2014], Bardi et al. [2017] and Hyde et al. [2015] instructed participants to detect whether the ball (the target content of the protagonist's true-/false-belief) was present or absent, or to passively view the movies, respectively. Thus, there was no way to guarantee that the TPJ activity observed in these studies does not reflect some level of explicit belief processing. Hyde et al. [2015] did, however, survey participants using a similar funneled debriefing questionnaire to this study, and excluded those participants that showed any awareness of the belief manipulation, suggesting that participants in this study at least did not engage in explicit belief processing.

A further distinguishing factor between our task and that used by Hyde et al. [2015] was the inclusion of a final "choice" phase, where on false-belief trials, some participants

saw the protagonist open the empty box (an expected behavior given their belief), whereas other participants saw them open the box with the ball (an unexpected behavior given their belief). Hyde et al. found no differences in right TPJ activity between these two false-belief groups, which was not surprising given the small sample size of each group ($n = 10$ and 5 , respectively). This manipulation of the protagonist's behavior could have changed the nature of the false-belief trials, however, whereby the participant now had to monitor not only their belief state and that of the protagonist but also the expected behavior of the protagonist. Indeed previous work has shown the right TPJ is sensitive to the pre-existing expectations we have about a protagonist's behavior and its impact on mental state evaluations [Saxe and Wexler, 2005]. The right TPJ response to false-belief trials in Hyde et al. [2015] could therefore have been enhanced by additional expectancy processes, which were not present in our past or current designs.

The current findings corroborate our previous work implicating the role of the STS and precuneus in implicit ToM processing, in which we observed greater activity in these brain areas for false-beliefs, relative to true-beliefs [Schneider et al., 2014a]. Here we show that left STS and precuneus can also distinguish false-belief scenarios from situations that involve no implicit beliefs (albeit this effect was only marginal in the left STS, $t(21) = 1.78$, $P = 0.090$), suggesting that implicit false beliefs elicit a greater mentalizing load than both baseline/congruent belief processing scenarios. These findings fit with the idea the STS is involved in processing mental states from another person's perspective, while the precuneus is involved in processing self-referential mental states [Abu-Akel and Shamay-Tsoory, 2011]. Support for the role of the precuneus in self-representation comes from Lou et al. [2004] who observed greater activity in the precuneus in response to self-related versus other-related character traits. On the other hand, the STS was more active for judgements about others, compared with judgements about oneself, implying that this brain region is involved in representing others. Together these findings suggest that the precuneus and STS might play a role in processing mental states that have been attributed to the self versus others; processes that are likely to be core components of mentalizing [Apperly and Butterfill, 2009; Schurz et al., 2015], in both its implicit and explicit forms.

The precuneus and STS/posterior TPJ have also been more generally associated with the interplay between the dorsal and ventral attentional systems in the brain [Corbetta and Shulman, 2002]. The dorsal attentional system is responsible for directing attention toward stimuli that are task-relevant, whereas the ventral attentional system monitors and filters available information to reduce it down to what is most important [Abu-Akel and Shamay-Tsoory, 2011]. The middle frontal gyrus is a candidate region that has also been implicated in interactions between the dorsal and ventral attentional systems [Fox et al., 2006], and indeed, we also found it to be responsive during the

current implicit ToM task. This brain region has been associated with specific explicit ToM tasks (e.g., false-belief vs photograph stories [Gobbini et al., 2007]) and, similar to the precuneus, is more active when people make self-evaluations relative to judgements about others [Schulte-Ruther et al., 2007]. Given the apparent overlap between the brain regions involved in attention and mentalizing processes (particularly in the TPJ; e.g., see Mitchell [2008]), it has been suggested that ToM tasks might recruit attentional processes that are necessary when shifting between focusing on the internal world (i.e., the self or our own perspective) and the external world (i.e., the perspectives and mental states of others). As this ability to shift between different perspectives is critical for inferring another person's mental state, whether it be with or without awareness, it seems reasonable to assume that interactions between the dorsal and ventral attentional systems could play a role in implicit ToM processes [Schneider et al., 2014a] or underpin a more general belief processing system in the brain.

Collectively, we have provided novel evidence for a general belief processing system that is responsive to both implicit and explicit ToM processes, and systems that do not completely overlap from ToM processing. Extending upon our earlier work [Schneider et al., 2014a], we find a more substantial degree of overlap between the implicit and explicit ToM systems, including the right TPJ, right STS, precuneus, and left middle frontal gyrus. These findings provide neurophysiological support for the prominent theoretical distinction between systems involved in implicit and explicit, domain-general (stimulus modality invariant) ToM [Apperly and Butterfill, 2009; Clements and Perner, 1994]. Moreover, our findings illustrate the importance of using comparisons of belief presence (i.e., false-belief vs no-belief conditions) as well as comparisons of belief content (i.e., false-belief vs true-belief conditions) when trying to discern the full extent of the implicit, and general-belief, ToM networks. These different belief scenarios (and the degree to which belief states are truly implicit or not) could modulate the level of attentional engagement and task-relevance of the protagonist's mental state, which might in turn affect whether a given brain region is active during the implicit ToM task. Our results expand our understanding of the implicit ToM system in the human brain, and the extent to which it overlaps with, and differs from, its explicit counterpart.

ACKNOWLEDGMENTS

We thank Aiman Al Najjar, Nicole Atcheson, Zoie E. Nott, and Ashley York for their assistance with data collection.

REFERENCES

Abu-Akel A, Shamay-Tsoory S (2011): Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia* 49: 2971–2984.

- 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. *J Cogn Neurosci* 21:1179–1192.
- Apperly IA, Butterfill SA (2009): Do humans have two systems to track beliefs and belief-like states? *Psychol Rev* 116:953–970.
- Asplund CL, Todd JJ, Snyder AP, Marois R (2010): A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. *Nat Neurosci* 13:507–512.
- Bardi L, Desmet C, Nijhof A, Wiersema JR, Brass M (2017): Brain activation for spontaneous and explicit false belief tasks overlaps: New fMRI evidence on belief processing and violation of expectation. *Soc Cogn Affect Neurosci* 12:391–400.
- Bargh JA, Chartrand TL (2009): The mind in the middle: A practical guide to priming and automaticity research. In: Reis HT & CM Judd, editors. *Handbook of Research Methods in Social and Personality Psychology*. New York: Cambridge University Press. pp 253–284.
- Baron-Cohen S, Leslie AM, Frith U (1985): Does the autistic child have a “theory of mind”? *Cognition* 21:37–46.
- Baron-Cohen S, Richler J, Bisarya D, Guranathan N, Wheelwright S (2003): The systemizing quotient: An investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. *Philos T Roy Soc B* 358:361–374.
- Baron-Cohen S, Wheelwright S (2004): The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *J Autism Dev Disord* 34:163–175.
- Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E (2001): The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J Autism Dev Disord* 31:5–17.
- Bayliss AP, Murphy E, Naughtin CK, Kritikos A, Schilbach L, Becker SI (2012): “Gaze leading”: Initiating simulated joint attention influences eye movements and choice behavior. *J Exp Psychol Gen* 142:72–92.
- Bayliss AP, Tipper SP (2006): Gaze cues evoke both spatial and object-centered shifts of attention. *Percept Psychophys* 68: 310–318.
- Brune M (2005): “Theory of mind” in schizophrenia: A review of the literature. *Schizophr Bull* 31:21–42.
- Butterfill SA, Apperly IA (2013): How to construct a minimal theory of mind. *Mind Lang* 28:606–637.
- Castelli F, Happé F, Frith U, Frith C (2000): Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* 12: 314–325.
- Clements WA, Perner J (1994): Implicit understanding of belief. *Cogn Dev* 9:377–395.
- Cohen AS, German TC (2009): Encoding of others' beliefs without overt instruction. *Cognition* 111:356–363.
- Corbetta M, Shulman GL (2002): Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* 3: 201–215.
- Dodell-Feder D, Koster-Hale J, Bedny M, Saxe R (2011): fMRI item analysis in a theory of mind task. *Neuroimage* 55:705–712.
- Dönel K, Schuwerk T, Meinhardt J, Sodian B, Hajak G, Sommer M (2012): Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *Neuroimage* 60:1652–1661.
- Fox MD, Corbetta M, Snyder AZ, Vincent JL, Raichle ME (2006): Spontaneous neuronal activity distinguishes human dorsal and

- ventral attention systems. *Proc Natl Acad Sci USA* 103: 10046–10051.
- Frith C, Frith U (2005): Theory of mind. *Curr Biol* 15:644–646.
- Gallagher HL, Happé F, Brunswick N, Fletcher PC, Frith U, Frith CD (2000): Reading the mind in cartoons and stories: An fMRI study of ‘theory of mind’ in verbal and nonverbal tasks. *Neuropsychologia* 38:11–21.
- Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, Haxby JV (2007): Two takes on the social brain: A comparison of theory of mind tasks. *J Cogn Neurosci* 19:1803–1814.
- Henry JD, Phillips LH, Ruffman T, Bailey PE (2013): A meta-analytic review of age differences in theory of mind. *Psychol Aging* 28:826–839.
- Hyde DC, Aparicio Betancourt M, Simon CE (2015): Human temporal-parietal junction spontaneously tracks others’ beliefs: A functional near-infrared spectroscopy study. *Hum Brain Mapp* 36:4831–4846.
- Iacoboni M, Lieberman MD, Knowlton BJ, Molnar-Szakacs I, Moritz M, Throop CJ, Fiske AP (2004): Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *Neuroimage* 21:1167–1173.
- Kovács ÁM, Kühn S, Gergely G, Csibra G, Brass M (2014): Are all beliefs equal? Implicit belief attributions recruiting core brain regions of theory of mind. *PLoS ONE* 9:e106558.
- Kovács ÁM, Téglás E, Endress AD (2010): The social sense: Susceptibility to others’ beliefs in human infants and adults. *Science* 330:1830–1834.
- Krall SC, Rottschy C, Oberwilling E, Bzdok D, Fox PT, Eickhoff SB, Fink GR, Konrad K (2015): The role of the right temporo-parietal junction in attention and social interaction as revealed by ALE meta-analysis. *Brain Struct Funct* 220:587–604.
- Lou HC, Luber B, Crupain M, Keenan JP, Nowak M, Kjaer TW, Sackeim HA, Lisanby SH (2004): Parietal cortex and representation of the mental self. *Proc Natl Acad Sci USA* 101:6827–6832.
- Low J, Watts J (2013): Attributing false beliefs about object identity reveals a signature blind spot in humans’ efficient mind-reading system. *Psychol Sci* 24:305–311.
- Maylor EA, Moulson JM, Muncer A-M, Taylor LA (2002): Does performance on theory of mind tasks decline in old age? *Brit J Dev Psychol* 93:465–485.
- Mitchell JP (2008): Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cereb Cortex* 18:262–271.
- Molenberghs P, Johnson H, Henry JD, Mattingley JB (2016): Understanding the minds of others: A neuroimaging meta-analysis. *Neurosci Biobehav Rev* 65:276–291.
- Moran JM, Young LL, Saxe R, Lee SM, O’Young D, Mavros PL, Gabrieli JD (2011): Impaired theory of mind for moral judgment in high-functioning autism. *Proc Natl Acad Sci USA* 108: 2688–2692.
- Onishi KH, Baillargeon R (2005): Do 15-month-old infants understand false beliefs? *Science* 308:255–258.
- Perner J, Lang B (1999): Development of theory of mind and executive control. *Trends Cogn Sci* 3:337–344.
- Phillips J, Ong DC, Surtees AD, Xin Y, Williams S, Saxe R, Frank MC (2015): A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychol Sci* 26: 1353–1367.
- Premack D, Woodruff G (1978): Does the chimpanzee have a theory of mind? *Behav Brain Sci* 1:515–526.
- Ramsey R, Hansen P, Apperly I, Samson D (2013): Seeing it my way or your way: Frontoparietal brain areas sustain viewpoint-independent perspective selection processes. *J Cogn Neurosci* 25:670–684.
- Razza RA, Blair C (2009): Associations among false-belief understanding, executive function, and social competence: A longitudinal analysis. *J Appl Dev Psychol* 30:332–343.
- Rice K, Redcay E (2016): Interaction matters: A perceived social partner alters the neural processing of human speech. *Neuroimage* 129:480–488.
- Ruffman T (2014): To belief or not belief: Children’s theory of mind. *Dev Rev* 34:265–293.
- Saxe R (2009): Theory of mind (neural basis). In Banks WP, editor. *Encyclopedia of Consciousness*, Vol. 2. Oxford, UK: Academic Press. pp 401–410.
- Saxe R, Kanwisher N (2003): People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind”. *Neuroimage* 19:1835–1842.
- Saxe R, Wexler A (2005): Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia* 43:1391–1399.
- Schneider D, Bayliss AP, Becker SI, Dux PE (2012a): Eye movements reveal sustained implicit processing of others’ mental states. *J Exp Psychol Gen* 141:433–438.
- Schneider D, Lam R, Bayliss AP, Dux PE (2012b): Cognitive load disrupts implicit theory-of-mind processing. *Psychol Sci* 23:842–847.
- Schneider D, Nott ZE, Dux PE (2014a): Task instructions and implicit theory of mind. *Cognition* 133:43–47.
- Schneider D, Slaughter VP, Bayliss AP, Dux PE (2013): A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition* 129:410–417.
- Schneider D, Slaughter VP, Dux PE (2015): What do we know about implicit false-belief processing. *Psychonomic Bulletin & Review* 22:1–22.
- Schneider D, Slaughter VP, Becker SI, Dux PE (2014b): Implicit false-belief processing in the human brain. *Neuroimage* 101:268–275.
- Schneider D, Slaughter VP, Dux PE (2017): Current evidence for automatic Theory of Mind processing in adults. *Cognition* 162:27–31.
- Schulte-Ruther M, Markowitsch HJ, Fink GR, Piefke M (2007): Mirror neuron and theory of mind mechanisms involved in face-to-face interactions: A functional magnetic resonance imaging approach to empathy. *J Cogn Neurosci* 19:1354–1372.
- Schurz M, Kronbichler M, Weissengruber S, Surtees A, Samson D, Perner J (2015): Clarifying the role of theory of mind areas during visual perspective taking: Issues of spontaneity and domain-specificity. *NeuroImage* 117:386–396.
- Schurz M, Radua J, Aichhorn M, Richlan F, Perner J (2014): Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neurosci Biobehav Rev* 42:9–34.
- Scott RM, Baillargeon R (2017): Early false-belief understanding. *Trends Cogn Sci* 21:237–249.
- Senju A, Southgate V, Snape C, Leonard M, Csibra G (2011): Do 18-month-olds really attribute mental states to others? A critical test. *Psychol Sci* 22:878–880.
- Senju A, Southgate V, White S, Frith U (2009): Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science* 325:883–885.
- Slaughter V, Peterson CC, Moore C (2013): I can talk you into it: Theory of mind and persuasion behavior in young children. *Dev Psychol* 49:227–231.
- Sommer M, Döhl K, Sodian B, Meinhardt J, Thoermer C, Hajak G (2007): Neural correlates of true and false belief reasoning. *Neuroimage* 35:1378–1384.
- Southgate V, Senju A, Csibra G (2007): Action anticipation through attribution of false belief by 2-year-olds. *Psychol Sci* 18:587–592.

- Surian L, Caldi S, Sperber D (2007): Attribution of beliefs by 13-month-old infants. *Psychol Sci* 18:580–586.
- Talairach G, Tourmoux P (1988): *Co-Planar Stereotaxic Atlas of the Human Brain*. New York: Thieme.
- van Overwalle F (2009): Social cognition and the brain: A meta-analysis. *Hum Brain Mapp* 30:829–858.
- Van Overwalle F, Vandekerckhove M (2013): Implicit and explicit social mentalizing: Dual processes driven by a shared neural network. *Front Hum Neurosci* 7:1–6.
- Wagner DD, Kelley WM, Heatherton TF (2011): Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. *Cereb Cortex* 21:2788–2796.
- Wellman HM, Cross D, Watson J (2001): Meta-analysis of theory-of-mind development: The truth about false belief. *Child Dev* 72:655–684.
- Wheatley T, Milleville SC, Martin A (2007): Understanding animate agents distinct roles for the social network and mirror system. *Psychol Sci* 18:469–474.
- Wimmer H, Perner J (1983): Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13:103–128.