Human Temporal-Parietal Junction Spontaneously Tracks Others' Beliefs: A Functional Near-Infrared Spectroscopy Study

Daniel C. Hyde,[1]* Mariana Aparicio Betancourt,[2] and Charline E. Simon[1]

[1]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, Illinois
[2]Neuroscience Program, Beckman Institute, University of Illinois Urbana-Champaign, Urbana, Illinois

Abstract: Humans have the unique capacity to actively reflect on the thoughts, beliefs, and knowledge of others, but do we also track mental states spontaneously when observing other people? We asked this question by monitoring brain activity in belief-sensitive cortex using functional near-infrared spectroscopy (fNIRS) during free-viewing of social videos. More specifically, we identified a portion of the right temporal-parietal junction (rTPJ) selective for mental state processing using an established, explicit theory of mind task, and then analyzed the brain response in that region of interest (ROI) during free-viewing of video clips involving people producing goal-directed actions. We found a significant increase in oxygenated hemoglobin concentration in our rTPJ ROI during free-viewing for all of our test videos. Activity in this region was further modulated by the extent to which the knowledge state, or beliefs, of the protagonist regarding the location of an object contrasted with the reality of where the object was hidden. Open-ended questioning suggested our participants were not explicitly focusing on belief states of the characters during free-viewing. Further analyses ruled out lower-level details of the video clips or general attentional differences between conditions as likely explanations for the results. As such, these results call into question the traditional characterization of theory of mind as a resource intensive, deliberate process, and, instead, support an emerging view of theory of mind as a foundation for, rather than the pinnacle of, human social cognition. *Hum Brain Mapp 00:000–000, 2015.* © 2015 Wiley Periodicals, Inc.

Key words: theory of mind; cognition; fNIRS; optical imaging; temporal-parietal junction

INTRODUCTION

Background

Successful human social interaction requires us to keep in mind the knowledge, or mental states, of others to interpret, predict, and plan behavior. The ability to think about others' thoughts and beliefs is referred to in the psychological literature as having a *theory of mind* [Premack and Woodruff, 1978]. Directly or indirectly cueing participants to think about the thoughts and knowledge of others activates a network of brain regions in the parietal, temporal, and prefrontal cortex [Koster-Hale and Saxe, 2013].

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

*Correspondence to: Daniel C. Hyde, Brain and Cognitive Development Lab, 621 Psychology Building, 603 East Daniel Street, Champaign, IL 61820. E-mail: dchyde@illinois.edu

Received for publication 29 January 2015; Revised 12 August 2015; Accepted 13 August 2015.

DOI: 10.1002/hbm.22953

Published online 00 Month 2015 in Wiley Online Library (wileyonlinelibrary.com).

However, to date, it is unclear whether theory of mind is engaged automatically in social situations with no direct or indirect prompting to do so. As a window into the nature of theory of mind, we ask whether the beliefs of others are tracked automatically by examining activity in belief-sensitive cortical regions of the theory-of-mind network during free-viewing of video clips involving two characters interacting with an object in a goal-directed manner. Additionally, we ask whether activity in this network is further modulated by the extent to which the clips convey that the beliefs of an actress are consistent or inconsistent with reality.

Traditionally, theory of mind has been characterized as a hallmark of human cognitive development, requiring substantial cognitive resources, not fully present in even the most sophisticated of non-human primates, showing a protracted human developmental timescale, and impaired in some individuals with developmental disabilities [Baron-Cohen et al., 1985; Wellman et al., 2001; Wimmer and Perner, 1983]. More specifically, proficiency at answering explicit questions about beliefs, and particularly false beliefs, demands executive resources like working memory and inhibition, does not emerge until 4 to 6 years, and is associated with continued brain development throughout late childhood [Apperly et al., 2008; Gweon et al., 2012; McKinnon and Moscovitch, 2007]. Other recent work, however, suggests that the traditional view of a deliberate, resource-intensive skill, requiring substantial experience and brain maturation before emerging may not be complete [see Baillargeon et al., 2010]. As evidence of this, adults can answer questions about the beliefs of characters in a story they just heard without being explicitly instructed to do so beforehand and the perspective or beliefs of others has been shown to implicitly influence one's own perspective [Apperly and Butterfill 2009; Cohen and German, 2009]. Furthermore, eliminating the general memory, inhibitory, linguistic, and/or motor response demands of explicit theory of mind reasoning tasks drastically reduces the age at which proficiency is observed in young children [Scott et al., 2012]. Finally, behavioral tasks measuring spontaneous looking patterns in infants suggests the ability to infer beliefs and even false beliefs of others may be engaged automatically and present from the first couple years of life [Baillargeon et al., 2010; Kovács et al., 2010]. For example, 18-month-old infants anticipate where an actress will reach by pre-emptively directing their gaze to a location where the actress (presumably) falsely believes an object is hidden compared with the actual location the object is hidden [Southgate et al., 2007; Senju et al., 2011]. Even younger infants display surprise, as measured by increased visual attending, when an actress makes a reach that runs contrary to her obtained knowledge about the location of an object [Kovács et al., 2010; Onishi and Baillargeon, 2005; Surian et al., 2007]. Together, these results call into question the validity of the traditional characterization and developmental timescale for theory of mind.

Despite substantial conjecture as to the interpretation of conflicting results in different paradigms and populations, it is still unclear from behavioral work how successes on implicit or spontaneous theory of mind tasks should be considered in relation to explicit theory of mind reasoning. Some suggest that emerging behavioral results from implicit or spontaneous theory of mind tasks support the idea of an automatic theory of mind, present early in development, continually working to anticipate behavior, and, thus, foundational to basic human social cognition [Baillargeon et al., 2010; see Cohen and German, 2009 for a review]. Under this view, what develops are the linguistic resources to appropriately describe others' thoughts and/or the information processing or executive resources to demonstrate theory of mind ability in more demanding tasks [He et al., 2012; Scott et al., 2012]. Others suggest that spontaneous theory of mind abilities arise from cognitive and perceptual abilities other than full-blown theory of mind, although the alternative characterization of these abilities varies substantially between accounts [Apperly and Butterfill, 2009; Frith and Frith, 2008; Low, 2010]. One common theme between accounts is that a more basic and general cognitive system could operate automatically, account for competencies in spontaneous tasks early in life or in adults without explicit instruction, and demand substantially less working memory, planning, and inhibition than that required for explicit tasks [Apperly and Butterfill, 2009; Low, 2010]. These accounts, however, diverge on whether or not this basic, or minimal, system involves actually representing the internal states of other people, like beliefs, or whether such competencies could be achieved through more general cognitive capacities such as the ability to learn and track spatiotemporal contingencies (e.g. between people and objects) [see Apperly and Butterfill, 2009 for a review). To date, then, behavioral methods have not been able to determine whether performance on spontaneous theory of mind tasks reflects engagement of common or distinct mechanisms from explicit theory of mind tasks.

To further understand the nature of theory of mind abilities, some have turned to measuring the brain. Functional neuroimaging studies have shown that theory of mind tasks reliably engage a network of brain regions, including the temporal-parietal junction (TPJ), medial prefrontal cortex (mPFC), and the superior temporal sulcus (STS) [see Koster-Hale and Saxe, 2013 for a review]. Of particular interest in this network has been the TPJ, as it appears to respond selectively to tasks involving reasoning about the knowledge or beliefs of others, with greater activation typically detected over the right hemisphere [e.g. Saxe and Kanwisher, 2003]. In contrast, mPFC and STS appear to play supportive roles, activating in a wide range of other types of tasks [e.g. Frith and Frith, 2006; Mitchell, 2009]. As such, TPJ has been characterized as a specialized brain region for mental state reasoning [Saxe, 2010; Saxe and Kanwisher, 2003].

Brain measures hold the potential to track cognition covertly [see examples in numerical cognition: Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Hyde and Spelke, 2009] and, thus, reveal the mechanisms engaged during spontaneous tasks. Most of our knowledge of the brain correlates of theory of mind to date, however, comes from explicit theory of mind tasks in which participants are told to directly focus on or answer questions about the beliefs of the characters in a story (read, heard, or seen) [Frith and Frith, 2000; Fletcher et al., 1995]. A smaller subset of neuroimaging studies do not ask directly about beliefs, but find engagement of the TPJ by using tasks or instructions that implicitly cue or even require reasoning about beliefs [Gweon et al., 2012; Saxe and Kanwisher, 2003; Mitchell et al., 2006; Sommer et al., 2007]. For example, Sommer et al. [2007] found engagement of the TPJ when asking participants to view and interpret the actions of a protagonist as expected or unexpected. Ma et al. [2011] found similar patterns of activation in the TPJ when participants were simply reading paragraphs or when reading with focus to answer questions about a character's traits. Other studies show activity in TPJ in a variety of social situations, such as moral reasoning, which very likely require belief inference about characters in the story for success [Castelli et al., 2000; Lombardo et al., 2010; Van Overwalle and Vandekerckhove, 2013]. While these implicit studies have informed our understanding of neural correlates of theory of mind, they fall short of demonstrating the neural correlates of spontaneous or automatic theory of mind because success in the particular experiment necessitates belief inference, even if this is implicitly cued. For example, an often touted study by Castelli et al. [2000] showing greater TPJ engagement while participants watched animations of geometric shapes that moved in a goal-directed and interactive manner compared with a random manner likely cued belief inference, as the researchers asked participants to interpret the actions of the shapes after every animation and provided instructions and practice examples of each type of animation prior to scanning. Heretofore we refer to such studies as implicit theory of mind experiments because they implicitly cue participants to engage in belief inference, or theory of mind reasoning, by way of elicited responses (verbal or non-verbal) or other demands of the task.

To our knowledge, only two studies to date have investigated the neural basis of spontaneous theory of mind, without instructions, elicited responses, or task demands that necessitate mental state reasoning [Kovács et al., 2014; Schneider et al., 2014] with conflicting results. One recent study used functional magnetic resonance imaging (fMRI) to ask whether regions of the mPFC and TPJ found to be active in previous explicit theory of mind studies were spontaneously sensitive to others' beliefs regarding the presence or absence of an object [Kovács et al., 2014]. To test this, brain activity was recorded while participants watched a series of animated video clips containing a cartoon character, a ball, and an occluder. All animations started by depicting a character watching a ball move behind an occluder. Under half of the conditions, the character was watching as the ball moved off-screen, after which the character left the scene likely believing the ball was gone. On the other half of the trials, the character left the scene likely thinking the ball was behind the occluder and then the ball moved off-screen. During a final outcome phase of the clips, the character returned to witness the occluder dropping to either reveal the ball or reveal nothing. To maintain attention, participants were simply asked to indicate whether or not the ball was present by pressing a button. Brain imaging results most notably revealed a significant increase in right TPJ activity during the belief formation phase of clips where the character was likely to have falsely believed the ball was present (e.g., last saw it behind the occluder, but it had later left the scene), compared with clips when the character was likely to have falsely believed the ball was gone (last saw it leave the scene, but the ball had returned after he had left the scene), and compared with both conditions when the character's belief regarding the location of the ball was consistent with reality (present or absent). Based on these results, the authors suggested that theory of mind can be engaged spontaneously, but only under the restricted conditions where beliefs represent positive content (e.g., the object is present), not negative content (e.g., the object is absent) [Kovács et al., 2014].

The conclusions of Kovács et al. contrast with those of a second recent study by Schneider et al. [2014], who conclude that belief-sensitive TPJ is not spontaneously engaged for representing the mental states of others. This study modified a well-known spontaneous behavioral paradigm [Southgate et al., 2007; also see Onishi and Baillargeon, 2005] in an attempt to covertly measure the neural mechanisms of spontaneous theory of mind with fMRI [Schneider et al., 2014]. To do this, researchers presented participants with video clips of two characters interacting with an object [e.g. Schneider et al., 2012; Southgate et al., 2007]. All video clips involved an actress observing a puppet move an object into one of two opaque boxes and the actress leaving the room. The two conditions differed in whether or not the actress left the room before or after the puppet moved the object a second time to the other box. Researchers compared the response across the whole brain during the final test video sequence (which was the same for both conditions) in which the puppet moved the object a third time back to the original location and the actress re-entered the room. The presumed knowledge or belief of the actress about the location of the hidden object differed between the two conditions depending on whether she had directly observed the second move of the object or not. Under the false belief condition, the actress's presumed knowledge about the location of the hidden object at the end of the test clip was inconsistent with its actual location. Under the true belief condition, the actress's presumed knowledge about the location of the hidden object at the end of the test clip was consistent with its actual

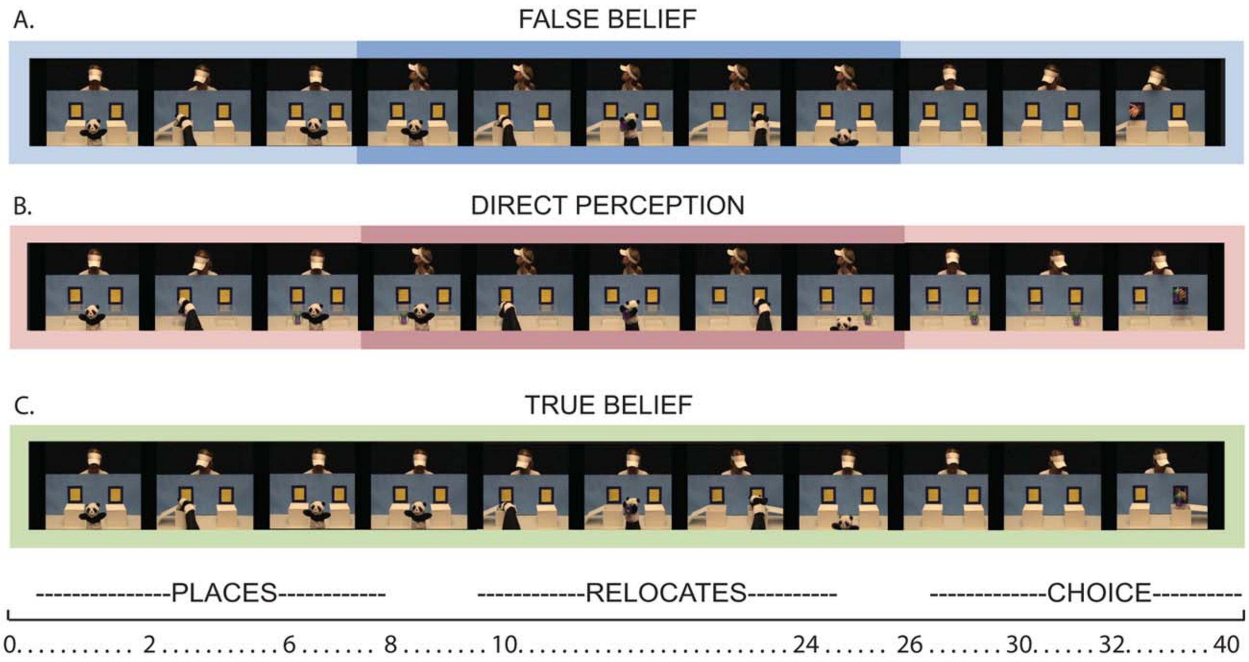

Figure 1.

Screen shots for events of interest during test video clips. Darker colored boxes around the center of the false belief and direct perception condition represent the period over which the actress was not looking at the puppet. The approximate time scale of events in seconds is placed below clips as a point of temporal reference (please refer to Supporting Information for more precise timing details).

condition. They predicted that the region(s) of the theory of mind network that are engaged spontaneously should show greater activation when the actress's belief was inconsistent with the location of the hidden object compared with when the actress's belief was consistent with reality. Using an explicit theory of mind localizer to define the theory of mind network, they observed that only a subset of the regions in the theory of mind network, left STS and posterior cingulate (PC), showed the predicted pattern (false belief > true belief) during the spontaneous theory of mind video clips. Based on these results, the authors claim that spontaneous or implicit theory of mind recruits a subset of nodes in the theory of mind network and this subset does not include the TPJ [Apperly and Butterfill, 2009; Clements and Perner, 1994]. Given the conflicting results of these two studies, it is unclear whether the neural correlates of spontaneous theory of mind have been appropriately characterized.

Current Study

Here we investigated the neural basis of spontaneous theory of mind and its relation to the explicit theory of mind reasoning network. To do so, we modified the spontaneous theory of mind paradigm used in developmental studies [Southgate et al., 2007; Senju et al., 2011; Onishi and Baillargeon, 2005] for use with measures of functional brain activity like a recent fMRI study [Schneider et al., 2014]. Our stimuli were tightly controlled for (presumed) higher-level knowledge of the characters, as well as lower-level movements (see Supporting Information for sample stimuli). More specifically, all stimulus video clips portrayed a puppet hiding an object from an actress in one of two boxes and then moving that object to the other box before exiting the screen (see Fig. 1). In one condition, the actress watched the puppet the entire time. As such, the actress's presumed belief was continually consistent with reality, or the object's true location (True Belief Condition, TB). In a second condition, the actress was distracted and looked off-screen as the puppet moved the object. As such, the actress likely held a belief about the location of the object that was inconsistent with the true location of the object (False Belief Condition, FB). Under a novel third condition, the actress also looked away, but the boxes were clear, allowing direct perception to the location of the object upon turning back around despite the momentary distraction (Direct Perception Condition, DP). Previous work has shown that even infants spontaneously take into account whether or not a person has visual access to an object in predicting actions, intentions, preferences, and goals [e.g., Luo and Baillargeon, 2007; Luo and Johnson, 2009] The demand to track mental states in the DP condition was comparable, then, to the demand when the actress had continual perception of the object in the true belief condition while matched for the exact motor movements of the false belief condition. While participants watched these video clips, we recorded brain activity over right parietal, temporal, and frontal regions. Participants completed an explicit theory of mind reasoning task afterwards, where they read stories about people and answered questions about beliefs or facts in the story, to independently localize belief-selective cortical regions of interest (ROI) for analysis of free-viewing data. Finally, we asked a group of participants (a subset of which also contributed to the final fNIRS dataset) open-ended questions about the passive-viewing stimuli to determine the extent to which explicit theory of mind reasoning strategies may have been deployed during the free-viewing tasks even though no instructions to do so were provided [following Bargh and Chartrand, 2000; Schneider et al., 2012, 2014].

We predicted that if the social context of the video alone, here agents interacting intentionally with each other and with objects, cues the brain to track the beliefs of the actress, then we would observe an increase in activity above baseline (rest) in belief-sensitive brain regions for all videos, an indication that theory of mind was engaged spontaneously from viewing the agents acting in the video. As further evidence that beliefs were being tracked, rather than some other lower-level aspect of the stimuli and following the predictions of Schneider et al. [2014], we predicted that activity in belief-sensitive cortical regions to free-viewing would vary as a function of condition. Specifically, we reasoned that more activity would be observed when the mental state or belief of the actress contrasted with reality (and another agent in the video-FB condition) compared with when the actress's beliefs accorded with reality (and another agent in the video-TB and DP conditions).

Functional Near-Infrared Spectroscopy

We measured the brain response using functional near-infrared spectroscopy (fNIRS). FNIRS is a noninvasive optical imaging method used to estimate levels of oxygenated and deoxygenated hemoglobin in brain tissue by monitoring changes in scattering of near-infrared light (here at 690 and 830 nm) from the scalp [for reviews see Boas and Franceschini, 2009; Obrig and Villringer, 2003]. Like fMRI, it relies on the neurovascular coupling to make claims about brain activity [Kleinschmidt et al., 1996; Obrig and Villringer, 2003; Strangman et al., 2002]. We chose to use fNIRS over other neuroimaging methods for several reasons. First, while the spatial resolution of fNIRS (~1–3 cm) is poorer than that observed in modern fMRI experiments (~1–3 mm) and largely restricted to the cortical surface, its resolution appeared to be sufficient in both size and depth for the localization of important belief-sensitive regions of the theory of mind brain network, including TPJ [Boas et al., 2004; Cui et al., 2011; Dodell-Feder et al., 2011; Saxe and Kanwisher, 2003]. Second, if the theory of mind network was engaged during free-viewing, we were interested to know what aspect of the stimuli evoked this engagement. Although both fMRI and fNIRS measure the intrinsically slow hemodynamic response, fNIRS has a significantly higher temporal sampling resolution (here 50 Hz) than fMRI (typically 0.5–1 Hz). In contrast to fMRI studies that typically focus on the average or peak response across an entire stimulus or block of stimuli [e.g. single peak as in Schneider et al., 2014], fNIRS allows for continual monitoring of activity across dynamic stimuli such as those to be used in our study. The temporal resolution of fNIRS would allow us to not only determine whether there were differences in the average or peak brain response, but also dynamically monitor changes in brain activity associated with different portions of our stimulus videos, not commonly done with fMRI, to determine when conditions might differ. Third, we were interested in ultimately developing a measure and method to test spontaneous theory of mind across development in both typical and atypical populations important to both theory and clinical application in theory of mind research. FNIRS does not require severely restricting motion of the subject like fMRI and, thus, can be (and has been) successfully applied across these populations of interest [Chaudhary et al., 2011; Hyde et al., 2010; Lloyd-Fox et al., 2013; Yucel et al., 2014].

MATERIALS AND METHODS

Participants

Adult participants were recruited through the University of Illinois psychology study pool and received course credit for their participation. Approval was obtained from the Institutional Review Board at the University of Illinois to conduct the study and all participants provided informed consent before any measurements were taken. The final fNIRS dataset consisted of 25 participants (13 females; *M* age = 19.5, SD age = 1.53). Nine others completed the study, but were eliminated from the data analysis after pre-processing because of poor signal quality and/or excessive artifacts during the experiment resulting in less than two artifact-free trials per experimental condition in each task (see Methods and Supporting Information for additional details). One additional participant was excluded after expressing knowledge of our theory of mind paradigm after the experiment.

Design and Procedure

The experiment was comprised of two main parts presented in a fixed order: a spontaneous theory of mind task [e.g. Schneider et al., 2014; Southgate et al., 2007] and an explicit theory of mind task [Dodell-Feder et al., 2011]. The entire testing session was limited to 1 hour.

Spontaneous theory of mind task

In the spontaneous theory of mind task, participants freely viewed silent video clips of a puppet interacting with an actress and objects, without any particular instruction, task, or cue (implicit or explicit) to focus on beliefs (see Fig. 1; see Supporting Information for sample videos). Participants were simply instructed at the beginning of the procedure to pay close attention to the video, as they would be asked questions afterwards. Although similar in content to behavioral studies of spontaneous theory of mind [Onishi and Baillargeon, 2005; Southgate et al., 2007], presentation parameters were modified for use with fNIRS.

Each run began with a black screen containing a fixation cross (12 s), followed by an introductory video clip (48 s) that presented a novel actress (waved at the camera), a novel object, and a puppet (see Supporting Information for examples). Introductory clips started with the new actress and a puppet waving at each other and then both waving at the camera. The rest of the introductory clips showed the puppet picking up the object and placing it in one of the two open boxes, after which the actress reached for and grabbed the object (once on each side). The actress continually watched, or had direct perceptual access to, the object in the open boxes during the entire video. As such, these videos were meant to introduce the participant to a new actress and demonstrate that the actress had a goal or disposition to obtain the object.

Three test video clips followed each introductory clip in each run, one from each experimental condition, presented in a random order (see sample clips in Supporting Information). Each test video clip was presented for 42 s and was both preceded and followed by a 12 s fixation screen. During the false belief condition test clips, the actress observed the puppet placing an object in one of the two opaque boxes. The actress then turned her head away from the puppet and the two boxes. While the actress was not looking, the puppet removed the object from its original box, placed it in the other box, and exited the scene. At the end of the false belief test clips the actress returned her gaze towards the table and boxes and reached into one of the boxes (see Fig. 1a). In the direct perception condition, the actress followed the same sequence of events except the boxes were clear and, thus, her returned gaze to the front allowed direct perceptual access to the new location of the hidden object (Fig. 1b). In the true belief condition, the actress never turned her head away, and, thus, was watching while the puppet relocated the object (Fig. 1c). In the true belief and direct perception clips for all participants, the actress always reached to the true location of the object and successfully obtained it. A subset of participants (n = 15) saw false belief test clips that ended in the actress also reaching for the true location and successfully obtaining the object; the remaining participants (n = 10) saw false belief test clips that ended in the actress reaching towards the original location of the objects, failing to find the object (as it had been moved). This between-subjects difference allowed us to also analyze the effects of the action performed by the actress relative to presumed knowledge/belief states of the actress on brain activity. All participants watched four total runs with three experimental test clips in each run for a total of 12 test trials (four per condition) lasting approximately 15 min. An independent eye-tracking study confirmed that, in fact, our stimuli elicited looking behavior consistent with spontaneous belief tracking as has been observed in other studies [Schneider et al., 2014; Southgate et al., 2007; see Supporting Information for study details]. Some participants were also subjected to open-ended questioning regarding the nature of the experiment as a method to gauge what participants were actually thinking/doing during the spontaneous task [cf. Schneider et al., 2012, 2014]. In particular, several questions were taken and modified from a questionnaire that has been used in previous eye-tracking and neuroimaging studies of spontaneous theory of mind [see Supporting Information; Bargh and Chartrand, 2000; Schneider et al., 2012, 2014]. Responses were coded for whether or not the participant used mental state terms (belief/believes, thought(s)/think(s), knowledge/know/know(s), and/or want(s)), as has been done in other studies [Schneider et al., 2012, 2014].

Explicit theory of mind reasoning task

During the explicit theory of mind task, participants read single-paragraph stories for 18 s and then answered a true or false question about the preceding story within 5 s. A total of 20 stories/questions were presented: 10 questions required belief reasoning about the character(s) in the story and 10 questions required factual reasoning about the state of events or objects in the story. Trials were separated by 12 s of rest/fixation. The entire task lasted approximately 12 min. The stories and questions have been used in fMRI studies of theory of mind to functionally localize the belief-selective cortical regions [Dodell-Feder et al., 2011] and have been made freely available to interested researchers (http://saxelab.mit.edu/superloc.php).

fNIRS Brain Measure

We measured the brain response using fNIRS [see Aslin and Mehler, 2005; Gervain et al., 2011 for a review of the technique]. Specifically, we used a TechEn (TechEn, Inc., Milford, MA) CW6 (continuous wave) NIRS system with four emitters (each emitter contained two light sources at 690 nm and 830 nm wavelengths) and eight light detectors to obtain measurements at a sampling rate of 50 Hz [Franceschini et al., 2003]. A custom-made head probe configured the four emitter and eight detector optodes in a geometric pattern with 3 cm spacing to cover a wide patch of scalp over temporal-parietal junction, as well as the posterior temporal and lateral frontal lobe (see Fig. 2). The custom headgear was developed from a modified adult size large EEG Electro-Cap (Electro-Cap International, Inc.,

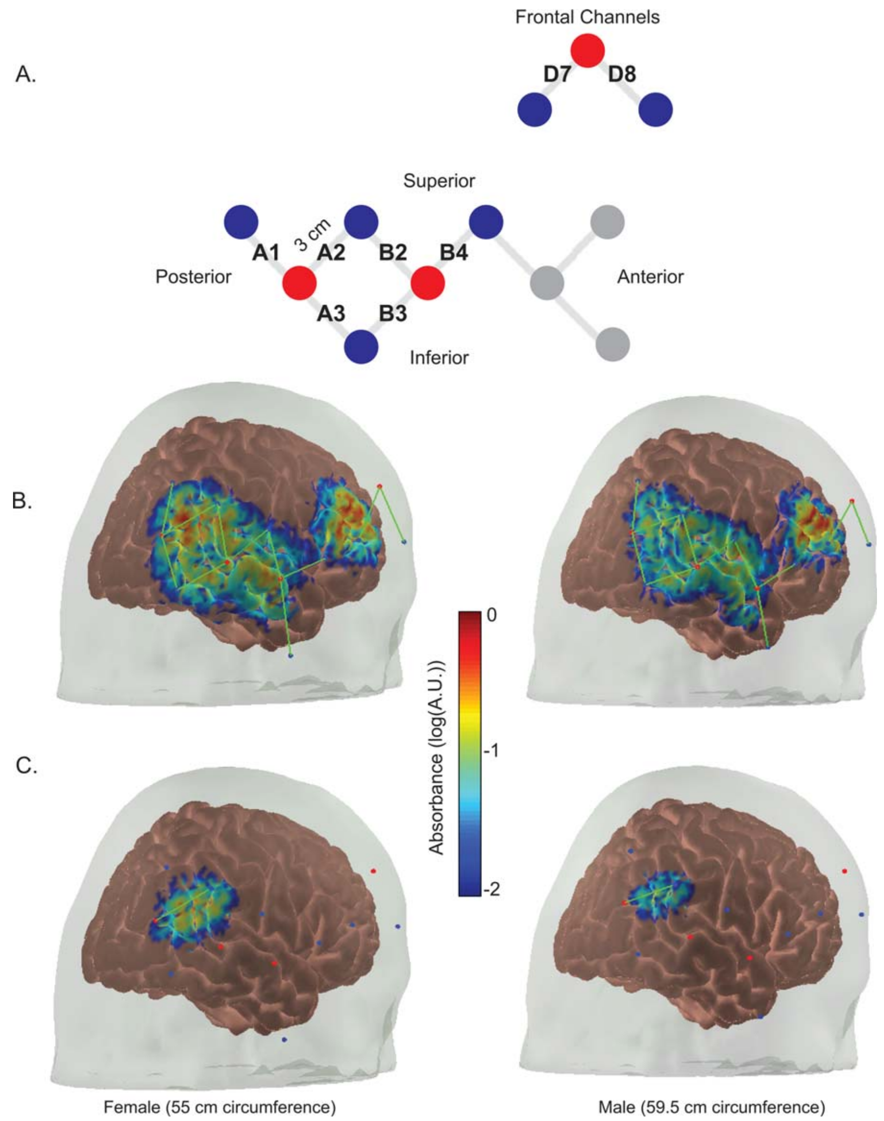

Figure 2.

Optical probe schematic and estimated cortical sensitivity map. **A**) Schematic of light source-detector channels used in our study. Red dots are light sources and blue dots are detectors. Bolded letter-number pairs in black font indicate channels used for analysis. Letters correspond to sources and numbers correspond to detectors. Gray dots in probe schematic were placeholders and not considered active in our analysis. **B**) Cortical sensitivity map for all source-detector pairs generated on a male and a female head. **C**) Cortical sensitivity map for TPJ region of interest (channel A-2) on a male and a female head.

Eaton, OH). Optodes were inserted into rubber grommets embedded within the cap to hold them against the scalp. Additional headbands were placed on over the head to further secure optodes against the scalp. Ten-foot optical fibers carried light to and from the head probe.

The head probe was placed on each participant's head relative to three scalp landmarks (10–20 coordinate Fpz, right preauricular point, left preauricular point), so that the probe was positioned in relatively the same way across participants. Initial measurements were taken for signal strength and quality. Any source-detector pair showing an attenuation value less than 60 db or greater than 140 db was manually adjusted by removing the optode from the grommet, physically displacing any impeding hair, and remeasuring the signal strength. Experimental recordings proceeded once all eight active channels of interest had signal strengths between 60 and 140 db (see Figure 2A).

We estimated the cortical sensitivity of each channel in our probe through simulations of photon migration in a realistic 3D head model using the mesh-based Monte

TABLE I. Average MNI coordinates for middle of active source-detector pairs

	Average MNI Position of Channel Pair Center ($x\ y\ z$)							
Head CC[a]	A1	A2	A3	B2	B3	B4	D7	D8
56.8	54 −61 32	54 −47 24	56 −55 7	60 −42 20	58 −42 1	71 −30 16	50 31 24	36 39 18

[a] Average head circumference value is in centimeters.

Carlo photon migration simulation algorithm implemented in AtlasViewer version 1.3.9 (implemented in Homer2 v 1.5.2) [for algorithmic details, see Fang, 2010]. The particular location of our probe was mapped to the surface of the default 3D head model of AtlasViewer through first digitizing the location of every optode in our probe relative to several scalp landmarks on five additional subjects and applying an affine transformation to localize the particular optode locations by corresponding the scalp landmarks in the subject-specific digitization with the default model [see Cooper et al., 2012; Custo et al., 2010]. More specifically, we digitized the location of each emitter and detector optode of our probe relative to 4 scalp landmarks (Nasion, Left preauricular, Right pre-auricular, Inion) and five 10 to 20 landmarks (Cz, F4, T8, P4, and P8) measured on the head of five additional participants (three females, two males; not in the original study) with a range of head sizes (54.5 cm, 55 cm, 56.5 cm, 58.5 cm, and 59.5 cm) using a Polhemus Patriot Digitizer (Polhemus Inc., Colchester, VT). Monte Carlo photon migration simulations of 1 million photons for each source and detector pair generated the sensitivity profile for each optode on the cortex (Fig. 2). Montreal Neurological Institute (MNI) coordinates corresponding to the center of sensitivity for each source-detector pair were also extracted based on the simulation data (see Table I).

fNIRS Data Preprocessing and Reduction

FNIRS data were pre-processed using freely available software [Homer2 v1.3; see Huppert et al., 2009 for a review). Raw light intensity data were first normalized and then converted to optical density. Several automated procedures were then used to objectively identify and eliminate problematic data. First, a principle components analysis was used to objectively identify and subsequently filter out systemic components of the data common to all channels, with the number of components removed constrained by removing no more than 80% of the total variance in the data of a given subject [see Cooper et al., 2102 for a review]. Second, an automatic individual channel pruning algorithm was applied to eliminate channels with poor data quality (too weak = mean light intensity over experiment < 0; too strong = mean light intensity over experiment >1 × 10^7; or if the signal to noise ratio was too low = mean intensity/standard deviation of intensity <3). Third, an automatic movement artifact detection algorithm was applied to the time course of data, where motion artifact was defined as a twofold change within a one second time window [see Huppert et al., 2009; or similar approaches Cooper et al., 2012; Scholkmann et al., 2010]. Five-second time windows of data were eliminated before and after any detected artifacts. Any epoch containing an eliminated time window between −5 and 30 s was rejected from further analysis for the explicit task and between −2 and 42 s for the spontaneous task. After channel pruning and artifact rejection, signals were converted to oxy- and deoxy-hemoglobin concentration using a modified Beer-Lambert Law [Obrig et al., 2000; Strangman et al., 2002]. Finally, data were high-pass filtered at 0.01 Hz to remove slow signal drift, low-pass filtered at 0.25 Hz to remove high frequency noise, (artifact-free trials) averaged by condition (−5 to 30 s for the explicit task and −2 to 42 s for the spontaneous task test trials), and baseline corrected to the average of the prestimulus rest portion of the epoch.[1]

fNIRS Data Analysis

Our analysis focused on the change in oxygenated hemoglobin (oxyHb) concentration, as previous studies have found it to be a more sensitive and reliable measure with fNIRS than change in deoxyHb concentration [Strangman et al., 2003]. For the explicit theory of mind task, we averaged oxyHb concentration separately over the time period in which participants were given to read the paragraphs (3–18 s) and respond to the question (18–26 s) for each channel for each subject. Conditions were compared using two-tailed, paired samples t-tests ($P < 0.05$) over the average oxyHb response extracted from each time period to determine which, if any, channels showed a significant difference between conditions.[2] Only one channel (A2, out of a possible 8; see Fig. 2) located over right TPJ showed the predicted response and, was therefore used as our independent ROI in the analysis of the data recorded during the free-viewing task.

[1] No differences between conditions were observed in the number of artifact free-trails retained for further data analysis in the explicit or spontaneous theory of mind tasks. Explicit theory of mind task: belief condition, $M = 9.24$ trials (SD = 1.16); fact condition, $M = 8.96$ trials (SD =1.34), $t_{(24)} = 1.16$, $P = 0.258$. Spontaneous theory of mind task/free viewing: direct perception, $M = 3.84$ (SD = 0.37); true belief, $M = 3.84$ (SD = 0.47); false belief, $M = 3.80$ (SD = 0.41). DP vs. TB: $t_{(24)} = 0.00$, $P = 1.0$; DP vs. FB: $t_{(24)} = 0.57$, $P = 0.574$; TB vs. FB: $t_{(24)} = 0.57$, $P = 0.574$.

[2] Correction for multiple comparisons was not employed in identification of ROIs given our main purpose was to liberally identify candidate ROIs for further analysis of the free-viewing data.

Three main types of analyses were conducted on the oxyHb response to passive viewing in the belief-sensitive TPJ ROI. First, we investigated whether our ROI showed a significant increase in activation at any point during the trial to any of our test conditions presented during passive-viewing. To do this, we found the average peak latency of the oxyHb response for each condition, extracted the mean oxyHb concentration for each participant over a symmetrical 2-s time window surrounding the peak (±1 s) to characterize the response for each condition, and statistically compared this response for each condition against the prestimulus baseline/zero (defined above) using a one-sample *t*-test.

Second, we compared the response across conditions for the predicted functional brain signature of belief processing irrespective of response timing following the primary peak analysis of Schneider et al. [2014]. More specifically, we compared the mean hemodynamic response surrounding the peak between conditions to see whether, irrespective of response timing, the false belief condition evoked a greater activation in our ROI than the true belief condition and the direct perception condition. Here, again, we used the mean oxyHb concentration over a 2-s symmetrical time window (±1 s) surrounding the peak latency for each condition (as above). This analysis allowed a comparison between conditions that was not biased to a particular time window, but considered the window of time with the greatest response independently for each condition.

Third, we took advantage of the temporal resolution of fNIRS (here 50 Hz) to compare conditions over each temporal sample (2,201 samples, −2 to 42 s) to determine when, if ever during our stimuli, conditions differed from each other. To do this, we first identified potential temporal clusters of data in the oxyHb time course where the predicted functional brain signature of belief processing was present. More specifically, a temporal cluster was defined as two or more consecutive time samples where the oxyHb response to the false belief condition was greater than oxyHb response to both the true belief and the direct perception conditions. Our precluster thresholding for identifying candidate clusters was set at $P < 0.05$, one-tailed, given our strong directional hypothesis and the fact that differences in the opposite direction would be treated equal to no differences (null hypothesis) [Kimmel, 1957; Ruxton and Neuhaüser, 2010]. Significance testing and cluster-based correction for multiple comparisons was carried out by permutation tests over the actual data [see Cohen, 2014 or Maris and Oostenvald, 2007]. Specifically, we randomly assigned condition labels to the actual data for each subject on each of 5000 permutations and then compared the actual observed cluster-size(s) in our data to the distribution of maximum cluster sizes obtained in the 5,000 random permutations of condition labels over the data. The same precluster thresholds ($P < 0.05$, one-tailed) that were used to obtain candidate clusters in the actual data were used to obtain the maximum clusters in the permuted data. As such, liberty in the precluster thresholding was equally prevalent in both the actual data and the permuted data. Significance, however, was determined by a cluster corrected significance (*P* value), calculated by dividing the number of max clusters (out of a potential 5,000) obtained through the permutations greater than the actual observed cluster in the data by the total number of permutations (5,000) [see Cohen, 2014 or Maris and Oostenvald, 2007]. Clusters were considered significant if the cluster corrected significance value (*P*) was less than 0.025. We used a more stringent cluster significance threshold of $P < 0.025$ (typically $P < 0.05$) to account for the fact that our testing (including permutations) only considered differences in one direction.

RESULTS

Behavioral Performance on Explicit Theory of Mind Reasoning Task

An analysis of behavioral performance in the explicit theory of mind reasoning task revealed similar accuracy and a small but reliable difference in reaction time across the belief and factual reasoning conditions (belief accuracy: M = 67%, SD = 20%; fact accuracy: M = 70%, SD = 16%; $t_{(24)} = -0.85$, P = 0.405; belief reaction time: M = 3,960 ms, SD = 537 ms; fact reaction time: M = 3,795 ms, SD = 486 ms; $t_{(24)}$ 2.37, P = 0.026; see Fig. 3).[3]

Explicit Theory of Mind Localizer Task

An analysis of the functional brain response over right parietal, temporal, and frontal lobe revealed only one channel over the right TPJ (see channel A2 in Fig. 2) that responded more when participants were reading and answering explicit questions about beliefs compared with when they were reading and answering questions about

[3] Accuracy in our fNIRS sample was considerably lower than previous fMRI studies. We believe two factors could have contributed to this. First, unlike previous studies employing fMRI where subjects are monetarily compensated well for their time, we sampled from an undergraduate study pool where participants were only receiving course credit. As such, our participants were likely to be less motivated than those in most other neuroimaging studies of theory of mind. Second, we required speeded responses (within 5 s), where answers outside of this time frame were counted as incorrect and assigned the maximum RT of 5,000 ms. Given the fact that the average response time was nearly 4 s and a reanalysis of the data excluding trials where participants failed to give a response within 5 s showed better performance (belief: M accuracy = 88%, SD = 12%; fact: M accuracy = 88%, SD = 9%), it is very likely that participants came up with the correct answers to a similar degree as others in previous studies but did not always indicate this within the 5 second time limit. We believe the combination of reduced motivation and the requirement of speeded responses, then, produced lower accuracy than is characteristically seen. Nonetheless, we have no reason to believe that such low performance would bias us to find spontaneous neural responses to the belief states of characters in our stimuli.

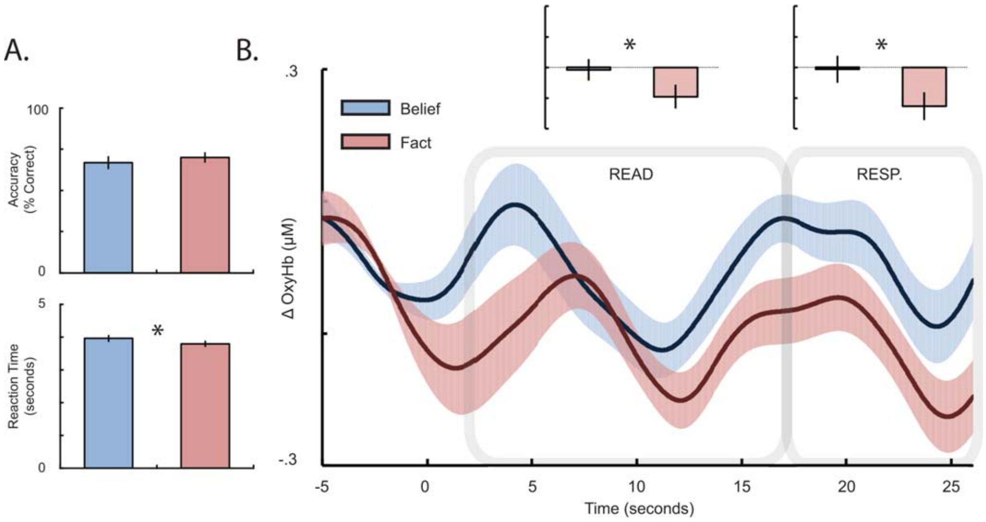

Figure 3.

Average behavioral and brain response to explicit theory of mind reasoning localizer task. **A**). Mean accuracy and reaction time for reasoning about beliefs and facts. **B**). Average time course of oxygenated hemoglobin (oxyHb) concentration change during explicit task over right temporal-parietal junction (channel A-2). Gray outlines represent the reading and response periods used for analysis. Bar graph represents mean response during time window of interest (scale ± 0.20 μM). Shaded regions in color around time course data and error bars on chart represent ± 1 standard error of the mean (SEM).

facts (reading: $t_{(24)} = 2.14$, $P = 0.043$; responding: $t_{(24)} = 2.36$, $P = 0.027$) (Fig. 3). A cortical sensitivity map generated using Monte Carlo photon migration simulation projected from digitized positioning of our probe on the scalp ($n = 5$) revealed that the sensitivity profile of this channel was comparable to rTPJ activity seen in previous fMRI studies of theory of mind (average MNI coordinates of center of sensitivity map: 54, −47, 24; Fig. 2; Table I). We defined this channel as our ROI for analysis of brain activity during the spontaneous theory of mind task. No other parietal, temporal, or pre-frontal channels from which we recorded showed significant differences between the belief and fact conditions during either the initial reading or response phase of the explicit theory of mind task (all $P > 0.09$).

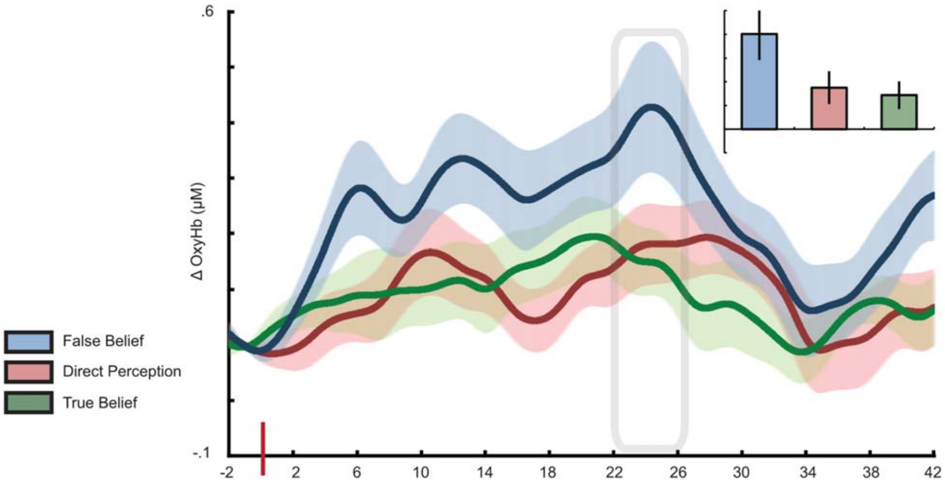

Figure 4.

Average time course of brain response during free-viewing task. **A**). Average oxyHb response of right TPJ ROI for each of the test conditions. Gray shaded box represents time window of significant differences in the predicted direction after permutation testing. Lighter colored regions around time course data and error bars on chart represent ± 1 SEM. Bar graph represents mean concentrations over time window of significant differences ($P < 0.025$ after temporal cluster correction) (bar graph scale: −.1/+.5 μM).

Brain Response During Spontaneous Theory of Mind Videos

We focused our analysis of the brain response during free-viewing within the belief-selective rTPJ ROI obtained from our independent explicit theory of mind localizer task, as no other data channels showed sensitivity to belief in the explicit task (free-viewing data was collected directly before explicit task data in our paradigm to avoid explicitly priming beliefs) using two complementary analytic approaches. First, we analyzed the mean brain response surrounding the peak of each condition to characterize the response irrespective of potential timing differences between conditions following the analysis of Schneider et al. [2014]. Second, using a novel data driven approach, we analyzed the entire time course for fixed temporal windows where conditions differed in the predicted direction.

An analysis of the mean hemodynamic response surrounding the peak for each condition revealed that activity significantly increased above baseline for all conditions (DP: peak = 27.82 s, $t_{(23)} = 2.94$, $P = 0.007$; TB: peak = 20.82 s, $t_{(23)} = 2.81$, $P = 0.010$; FB: peak = 24.36 s, $t_{(23)} = 3.62$, $P = 0.001$)[4], suggesting that free-viewing evoked an increase in oxyHb concentration in the belief-selective right temporal-parietal ROI at some point during our stimuli for all test video types (Fig. 4). Further comparison of the mean hemodynamic response surrounding the peak between conditions revealed that the false belief condition produced a larger response than the true belief condition ($t_{(23)} = -2.24$, $P = 0.035$) and a marginally greater response compared with the direct perception condition ($t_{(23)} = -2.09$, $P = 0.048$); no differences were observed between the direction perception condition and the true belief condition ($t_{(23)} = -0.04$, $P > 0.96$). The analysis of the spontaneous brain response to our stimuli independent of response timing, then, was consistent with the predicted pattern of results, as the response to false belief was greater than that to true belief and was no different from that to the direct perception condition. As such, it is possible that the response was sensitive to mental states of the actress. However, given the fact that this analysis did not consider the timing of the response, it is also possible that lower-level stimulus differences drove the observed pattern of differences.

A more careful analysis of the temporal dynamics of brain activity within the TPJ ROI during passive viewing revealed a robust period of sustained differences between conditions further indicative of the predicted belief-processing signature. In particular, a comparison of conditions across the entire time course of activity revealed several candidate periods of sustained differences (temporal data clusters) in the predicted direction, with the false belief condition producing a larger brain response than the direct perception condition and the true belief condition (three candidate clusters: Cluster 1, 105 samples, 4.40–6.48 s; Cluster 2, 165 samples, 11.18–14.46 s; Cluster 3, 214 samples, 22.22–26.48 s). Permutation testing showed that one of these clusters between 22.22 and 26.48 s was of sufficient size and strength to be considered statistically significant after temporal cluster correction for multiple comparison ($P < 0.025$) and, thus, not likely to have been obtained by chance (Cluster 3, $P = 0.019$; other non-significant clusters: Cluster 2, $P = 0.032$; Cluster 1, $P = 0.062$) (Fig. 4).[5] Considering the delay in the hemodynamic response is typically between 1 and 4 s, this significant temporal cluster corresponded to the end of the relocation portion of the false belief test clip when the puppet had surreptitiously moved the object to a new location while the actress was not looking and in anticipation of the actress facing forward again. In other words, this was the period during which the actress's belief about the location of the hidden object became inconsistent with actual location of the hidden object. Importantly, the puppet performed the exact same actions during this time frame in all conditions and the actress performed the exact same actions in the false belief condition as in the direct perception condition during this time frame, suggesting idiosyncratic low-level differences between movements of the puppet and/or actress in the false belief condition could not explain these differences. A more traditional fNIRS average response analysis over the three major time periods of the test video clips showed a similar pattern of results.[6]

Tests of Alternative Accounts of the Results

One difference between the false belief condition and the other conditions, however, was that a subset of the participants ($n = 10$) in the false belief condition saw the

[4] One of the 25 participants had a bad ROI channel on the spontaneous task (but otherwise good data). As such, the correct degrees of freedom were 22 instead of 23 for between-subjects *t*-tests, 23 instead of 24 for one-sample *t*-tests, and 23 instead of 24 for paired samples *t*-tests on the spontaneous task fNIRS data.

[5] As a purely exploratory measure to guide future studies, we searched the other non-ROI channels for the predicted pattern of results (FB > TB and FB > DP) using the same temporal cluster search parameters and permutation method for significance testing as was used within our ROI. This analysis revealed the predicted pattern of response in two non-belief selective channels surrounding our ROI (A1, 181 samples from 22.46 to 26.06 s, $P = 0.025$; B2, 185 samples from 3.10 to 6.78 s, $P = 0.024$, based on 5,000 randomly permutations of condition labels, see Figure 2a for relative locations to ROI channel A2).

[6] Three time periods were analyzed, corresponding to when the puppet initially placed the object (place: 3–10.5 s.), when the puppet relocated the object (relocate: 10.5–26.5 s.), and when the actress reached for the object (reach: 26.5–42 s) (see Fig. 1). Differences in the predicted direction (FB > TB and FB > DP) were only observed during the period in which the puppet was relocating the object (relocate: FB > DP, $t_{(23)} = 2.67$, $P = 0.014$; FB > TB, $t_{(23)} = 1.86$, $P = 0.076$*, *only marginally significant). No differences were observed at any other time window between conditions (all other P's > 0.09).

actress reach towards the old location at the end of the clip (expected outcome-where she would have likely thought the object was), while the rest of the participants in this condition saw the actress reach towards the actual location (unexpected outcome-where she presumably would not have known the object was hidden). No differences, however, were observed in our ROI during the reaching portion of the FB test video clip (26.5–42 s) between participants that saw the actress reach for the old location and participants that saw the actress reach for the new location ($t_{(22)} = -1.42$, $P = 0.170$). Further analysis also failed to reveal differences in the response between-subjects seeing different FB outcomes during the earlier portion of the stimuli that evoked differences between the FB condition and the other conditions (cluster 3: $t_{(22)} = 0.44$, $P > 0.66$). Most broadly, no significant differences in the ROI brain response were observed between any of the conditions during the reach period of the test video clips (all Ps > 0.11). Together these analyses suggest that this portion of rTPJ was not sensitive to lower-level differences in the reaching action, higher-level associations between the particular reaching actions and their outcomes (i.e., success or failure/expected vs. unexpected), or to whether the actress's reaching action was consistent or inconsistent with her beliefs or knowledge. The observed response in this ROI, then, appears to be sensitive to the knowledge or beliefs of others independently of and encoded from information other than the reaching action itself. This, of course, does not preclude the possibility that other brain regions would not be sensitive to such factors.

The observed functional response pattern (FB > TB and FB > DP) also served to rule out several alternative accounts of brain activity patterns. If the difference between conditions resulted from the object being placed in a clear versus opaque box, and, thus, being occluded under two conditions (FB and TB) but not in the other (DP), then greater activity should have been seen in the false belief and the true belief conditions relative to the direct perception condition. Similarly, if the difference between conditions was due purely to differences in eye-gaze of the actress between conditions, then the true belief condition should have been different than both the direct perception condition and the false belief condition, both of which involved averted eye gaze. However, at no time point in the data (2,201 points from −2 to 42 s), even before precluster thresholding, were significant differences observed between the true belief and direct perception conditions (all uncorrected Ps > 0.12). As such, neither of these alternative accounts holds and, thus, the results are not likely due to such lower-level differences between videos.

Finally, it is possible that our ROI was more generally sensitive to attentional differences between conditions, rather than beliefs, and the false belief condition produced a greater brain response because it was more interesting or elicited more attention than the other conditions. While this alternative is continually debated and notoriously hard to completely rule out, we attempted to test it by comparing the initial brain response to introductory video onset with the response to false belief test videos. For several reasons, the response to introductory clips is likely to have been more attentional engaging, in a general sense, than the false belief clips. First, introductory clips began with the presentation of a new actress while test clips involved the same actress that had been on the screen in prior introductory and test clips. Second, introductory clips involved the actress and the puppet directly interacting, including engaging and gesturing with each other and with the camera, whereas test clips only involved the actress and the puppet interacting with the object. If our ROI were generally sensitive to attentional differences between conditions, rather than specific to beliefs, it would then likely respond more to the introductory clips than the false belief condition. A direct comparison of the hemodynamic response between introductory clips and the false belief clips (using the same peak-picking procedure outlined for the peak analysis of conditions above) revealed that the brain response to the false belief condition was greater than the brain response than the introductory video clips (peak 13.78 s, 12.78–14.78 s) in our ROI ($t_{(23)} = -2.51$, $P = 0.019$), suggesting that the response of the rTPJ to the false belief condition was not likely to be driven solely by greater attention to the false belief condition than the other test conditions.

Behavioral Descriptions of Spontaneous Theory of Mind Videos

To determine the extent to which participants might have been engaged in explicit reasoning about thoughts or beliefs of the characters in our stimuli during the free-viewing task despite any instructions to do so, we asked 34 participants (15 of the 34 participants surveyed were also included in the final fNIRS dataset) open-ended questions regarding the nature of the experiment [Bargh and Chartrand, 2000; Schneider et al., 2012, 2014]. Participants overwhelmingly gave concrete descriptions of the actions and events as they unfolded in the video (e.g. a puppet appeared, the puppet moved the ball, a person reached for the ball, etc.) in their response to the questions. Only 1 of the 34 (2.9%) participants surveyed mentioned target mental state terms in their responses.[7] Based on these data, it appears that most participants were not likely to have been explicitly focusing on the mental states of the characters or engaged in explicit theory of mind reasoning during the free-viewing task.

[7] The subject that did mention beliefs also demonstrated explicit knowledge of the implicit theory of mind literature from a developmental psychology class and was not included in the final fNIRS dataset.

DISCUSSION AND CONCLUSIONS

Behavioral and brain imaging research on theory of mind has relied largely on tasks that implicitly or explicitly necessitate belief inference. From this work, the automaticity of theory of mind is unclear. In the few studies that truly measure spontaneous or automatic theory of mind abilities, such as visual attention studies with young infants or adults, it is unclear whether patterns of behavior reflect actual theory of mind engagement or arise from other general cognitive or social mechanisms. We attempted to overcome this limitation in our study by identifying belief-selective cortical mechanisms in an explicit task and then monitoring brain activity in those regions during a typical spontaneous theory of mind task. We hypothesized that such measures would allow us to covertly investigate whether or not the brain mechanisms of theory of mind can also be spontaneously engaged in the absence of task demands or elicited responses that necessitate belief inference. We observed a significant increase in activity during all of our free-viewing test conditions in the same right TPJ data channel independently identified as belief-selective in the explicit task. Photon simulation confirmed that this channel was spatially consistent with TPJ activity routinely observed in other studies as selective for theory of mind reasoning [Saxe and Kanwisher, 2003; Dodell-Feder et al., 2011].

The temporal resolution of our fNIRS method, combined with a sample-by-sample analysis, further allowed us to see that activity in the TPJ ROI was modulated by the extent to which the knowledge state, or beliefs, of a person regarding the location of an object contrasted with the reality of where the object was hidden. Furthermore, significant differences between conditions arose only during the end of a relocation period in the false belief condition when the puppet had surreptitiously moved the object from one opaque box to another.[8] This was the only period over all the test video clips when the reality of the location of the object came in conflict with the actress's presumed belief about the location of the object (and where there would be no direct perceptual access to location upon turning around). For the true belief condition, the actress's beliefs were continually consistent with the actual location of the object and for the direct perception condition the actress's beliefs would likely be updated to the actual location immediately upon turning around. We speculate that increased activity in belief-selective rTPJ for the false belief condition relative to the other conditions results from the demands on TPJ of holding at least two inconsistent belief states in mind in the false belief condition compared with the other conditions where belief states between the actress, the puppet, and reality were likely consistent with one another given the possibility of direct perceptual access. This functional response pattern, a noted neural signature of belief inference in some studies [Aichhorn et al., 2009; Schneider et al., 2014; Sommer et al., 2007], provides additional evidence that actual mental states were being tracked by the rTPJ during free-viewing.

It is not likely that the observed functional response pattern arose as a result of lower-level stimulus differences between conditions, as differences in the brain response did not pattern with predictions based on such lower-level stimulus differences. For example, the false belief condition differed from the true belief condition, but was equated with the direct perception condition, in actress movement and the false belief condition differed from the direct perception condition, but was equated with the true belief condition, in object occlusion (DP had transparent boxes), yet a greater TPJ response was observed to the false belief condition compared with both the true belief and the direction perception condition and no differences were observed in TPJ activity between the true belief and direct perception conditions. All stimuli were equated for puppet movement. Additional analyses supported the idea that the difference between the false belief condition and other conditions was not solely driven by differences in general attention between conditions. The combination of spatial overlap in the mechanisms of explicit and free-viewing tasks, the functional response pattern differences between conditions, the stimulus controls, and the additional analyses, together provide strong evidence that actual theory of mind was engaged during free-viewing. Thus, it appears that adults were automatically tracking the knowledge or beliefs of the actress in an on-line fashion despite no instruction to do so.

Two recent fMRI studies of spontaneous theory of mind also showed some overlap in cortical regions for explicit and spontaneous theory of mind [Kovács et al., 2014; Schneider et al., 2014]. One study reported some evidence that TPJ was involved in spontaneous theory of mind [Kovács et al., 2014]. Another study, however, reported no differences between their true belief and false belief conditions in the TPJ, claiming that TPJ was not part of the theory of mind network involved in automatic belief tracking [Schneider et al., 2014]. Using a comparable analysis to that of Schneider et al., we found differences in the TPJ brain response between false belief and true belief

[8] Although significant differences arose only at the end of the relocation period, marginal differences were observed earlier in the stimuli (e.g. ~11–14 s). Given that the lag in the hemodynamic response can be between 1 and 4 s, it is possible that differences in the brain were initiated on some trials shortly after the actress turned her head and in anticipation of the puppet initiating the object relocation, but were only actually present in the average brain response during the puppet movement itself. If this were the case, participants may have come to anticipate the events of the false belief based on their experience with preceding clips of false belief that always uniquely contained both the actress turning away and opaque boxes. Future studies should test this possibility directly by tracking the change in the hemodynamic response latency across repeated trials involving false belief.

conditions irrespective of response timing. The temporal resolution of fNIRS allowed us to further analyze the entire time course with respect to moment-by-moment conceptual changes in the stimuli. This analysis yielded robust differences in the predicted direction between conditions, but only at the portion of the stimuli when protagonist's presumed belief about the location of the hidden object became inconsistent (or false) with regards to the actual location. In contrast to the conclusions drawn from Schneider et al. [2014] and in support of the conclusions of Kovács et al. [2014], then, our data suggest that TPJ is involved in spontaneous theory of mind reasoning. However, it might be the case that TPJ involvement is most evident in the brain response when timing of the response is taken into account.

Implications for Cognitive Theory

These results have at least two theoretical implications regarding the nature of theory of mind. First, our results provide some of the strongest evidence to date that actual theory of mind can be engaged automatically and, thus, does not always require substantial executive resources, deliberate focus, or indirect cueing to be deployed. While previous behavioral and eye-tracking studies patterns have suggested this possibility, they have been unable to definitively determine the mechanisms that drive such behaviors [e.g. Kovács et al., 2010; Onishi and Baillargeon, 2005; Senju et al., 2011; Southgate et al., 2007; Surian et al., 2007]. That is, it is unclear if such indicative behavior arises from actual spontaneous belief tracking or by some other means. We provide evidence that one central component of the theory of mind network, the right TPJ, is engaged during free-viewing of typical spontaneous theory of mind stimuli and that functional response patterns strongly suggest that this activity is tracking beliefs without explicit instruction or reported awareness of doing so. We believe this provides evidence that theory of mind can be automatically, as well as explicitly engaged. While our approach narrowed the focus of analysis to brain region(s) showing selectivity for mental states during an explicit task, future work should investigate more general-purpose nodes of the theory of mind brain network, such as STS and PFC, during spontaneous theory of mind tasks.

Second, and relatedly, our results have implications for understanding the relationship between explicit and spontaneous theory of mind abilities. Some accounts of previous work resolve conflicts between behavioral results on spontaneous and explicit/elicited response theory of mind tasks by proposing distinct systems (or sub-systems) for spontaneous and explicit theory of mind, where each system represents qualitatively different conceptual content. Under such views, the automatic system is usually attributed with less conceptually rich content than the explicit system [see Apperly and Butterfill, 2009 for a review]. In contrast, others have proposed that a single, largely continuous, cognitive system operates in both explicit and spontaneous situations over the lifespan, where differences in executive and linguistic demands between tasks mask this continuity [He et al., 2012; Scott et al., 2012]. Still others present a hybrid-view where spontaneous theory of mind allows rich mental state representations but only of specific contents, such as the presence but not the absence of an object [Kovács et al., 2014]. While this debate is certainly not completely resolved by our results, we provide evidence that the automatic system overlaps, at least partially, in brain mechanism and in richness of content with the explicit system in adults. Furthermore, patterns of activity in this region did not appear to track other differences between conditions such as differences in movement, actions, or correspondences between actions and objects, often touted as lower-level factors underlying a minimal theory of mind system and differentiating it from a more sophisticated explicit system [e.g. Apperly and Butterfill, 2009; Clements and Perner, 1994]. Our paradigm did not directly manipulate the presence versus absence of the target object, and, as such, we cannot directly address whether the spontaneous theory of mind is selective for positive beliefs [Kovács et al., 2014]. Nevertheless, our data are consistent with the idea that reasoning about the mental states of others is a foundational cognitive process, which can be automatically engaged in the mind in situations potentially relevant to interpreting past or current behavior and predicting future behavior.

Future work is needed to determine whether such results hold earlier in development and across clinical populations and, as such, our study may have future methodological implications for developmental study and assessment of theory of mind. For example, individuals with autism have traditionally been characterized as impaired in theory of mind and this impairment has been cited as one source of broader social deficits commonly observed with autism spectrum disorders [Baron-Cohen et al., 1985]. However, evidence for impairment in theory of mind is largely derived from theory of mind tests that rely heavily on linguistic and executive processes, processes also commonly impaired in autism [Charman et al., 2003; Robinson et al., 2009]. Furthermore, theory of mind tasks that have been used often do not explicitly mention belief, but cue responses that require belief inferences, such as interpretation or response to an expected or unexpected outcome [see Senju, 2012 for a review]. Our free-viewing fNIRS method potentially provides a novel avenue for concurrently measuring theory of mind abilities and functional development of the social brain in non-verbal individuals and/or individuals with language impairment such as some with autism spectrum disorders, where traditional theory of mind tasks eliciting a verbal or non-verbal response may not be possible and the physical constraints of functional MRI are often challenging. More broadly, and in direct contrast to established fMRI techniques, our method could be applied to typically developing populations of virtually all ages from early infancy to late adulthood to compare and contrast socio-cognitive brain development across the entire lifespan.

ACKNOWLEDGMENTS

The authors thank Renée Baillargeon for her helpful comments on an early draft of the manuscript. The authors declare no known conflicts of interest.

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. J Cognit Neurosci 21:1179–1192.

Apperly IA, Back E, Samson D, France L (2008): The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. Cognition 106:1093–1108.

Apperly IA, Butterfill SA (2009): Do humans have two systems to track beliefs and belief-like states? Psychol Rev 116:953.

Aslin RN, Mehler J (2005): Near-infrared spectroscopy for functional studies of brain activity in human infants: Promise, prospects, and challenges. J Biomed Opt 10:011009–0110093.

Baillargeon R, Scott RM, He Z (2010): False-belief understanding in infants. Trends Cognit Sci 14:110–118.

Bargh JA, Chartrand TL (2000): The mind in the middle: A practical guide to priming and automaticity research. In: Reis HT, Judd CM editors. Handbook of Research Methods in Social and Personality Psychology. New York: Cambridge University Press. pp 253–285.

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

Boas DA, Dale AM, Franceschini MA (2004): Diffuse optical imaging of brain activation: Approaches to optimizing image sensitivity, resolution, and accuracy. Neuroimage 23:S275–S288.

Boas D, Franceschini MA (2009): Near infrared imaging. Scholarpedia 4:6997.

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.

Charman T, Drew A, Baird C, Baird G (2003): Measuring early language development in preschool children with autism spectrum disorder using the MacArthur Communicative Development Inventory (Infant Form). J Child Lang 30:213–236.

Chaudhary U, Hall M, Gutierrez A, Messinger D, Rey G, Godavarty A (2011): Joint attention studies in normal and autistic children using NIRS. *Proc. SPIE* 7883, Photonic Therapeutics and Diagnostics VII, 788348. doi:10.1117/12.874360

Clements WA, Perner J (1994): Implicit understanding of belief. Cognit Dev 9:377–395.

Cohen AS, German TC (2009): Encoding of others' beliefs without overt instruction. Cognition 111:356–363.

Cohen AS, German TC (2010): A reaction time advantage for calculating beliefs over public representations signals domain specificity for 'theory of mind'. Cognition 115:417–425.

Cohen MX (2014): Analyzing Neural Time Series Data: Theory and Practice. Cambridge, MA: MIT Press.

Cooper RJ, Caffini M, Dubb J, Fang Q, Custo A, Tsuzuki D, Boas DA (2012): Validating atlas-guided DOT: A comparison of diffuse optical tomography informed by atlas and subject-specific anatomies. Neuroimage 62:1999–2006.

Cooper RJ, Selb J, Gagnon L, Phillip D, Schytz HW, Iversen HK, Boas DA (2012): A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy. Front Neurosci 6:1–10.

Cui X, Bray S, Bryant DM, Glover GH, Reiss AL (2011): A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. Neuroimage 54:2808–2821.

Custo A, Boas DA, Tsuzuki D, Dan I, Mesquita R, Fischl B, Wells IIIW (2010): Anatomical atlas-guided diffuse optical tomography of brain activation. Neuroimage 49:561–567.

Dodell-Feder D, Koster-Hale J, Bedny M, Saxe R (2011): fMRI item analysis in a theory of mind task. Neuroimage 55:705–712.

Fang Q (2010): Mesh-based Monte Carlo method using fast ray-tracing in Plucker coordinates. Biomed Opt Expr 1:165–175.

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RS, Frith CD (1995): Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. Cognition 57:109–128.

Franceschini MA, Fantini S, Thompson JH, Culver JP, Boas DA (2003): Hemodynamic evoked response of the sensorimotor cortex measured noninvasively with near-infrared optical imaging. Psychophysiology 40:548–560.

Frith CD, Frith U (2000): The physiological basis of theory of mind: Functional neuroimaging studies. In: Baron-Cohen S, Tager-Flusberg H, Cohen DJ, editors. Understanding Other Minds: Perspective from Developmental Social Neuroscience. New York: Oxford University Press. pp 334–356.

Frith CD, Frith U (2006): The neural basis of mentalizing. Neuron 50:531–534.

Frith CD, Frith U (2008): Implicit and explicit processes in social cognition. Neuron 60:503–510.

Gervain J, Mehler J, Werker JF, Nelson CA, Csibra G, Lloyd-Fox S, Aslin RN (2011): Near-infrared spectroscopy: A report from the McDonnell infant methodology consortium. Dev Cogn Neurosci 1:22–46.

Gweon H, Dodell-Feder D, Bedny M, Saxe R (2012): Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. Child Dev 83:1853–1868.

He Z, Bolz M, Baillargeon R (2012): 2.5-year-olds succeed at a verbal anticipatory-looking false-belief task. Brit J Dev Psychol 30:14–29.

Huppert TJ, Diamond SG, Franceschini MA, Boas DA (2009): HomER: A review of time-series analysis methods for near-infrared spectroscopy of the brain. Appl Opt 48:D280–D298.

Hyde DC, Boas DA, Blair C, Carey S (2010): Near-infrared spectroscopy shows right parietal specialization for number in pre-verbal infants. Neuroimage 53:647–652.

Hyde DC, Spelke ES (2009): All numbers are not equal: An electrophysiological investigation of small and large number representations. J Cognit Neurosci 21:1039–1053.

Kimmel HD (1957): Three criteria for the use of one-tailed tests. Psychol Bull 54:351–353.

Kleinschmidt A, Obrig H, Requardt M, Merboldt KD, Dirnagl U, Villringer A, Frahm J (1996): Simultaneous recording of cerebral blood oxygenation changes during human brain activation by magnetic resonance imaging and near-infrared spectroscopy. J Cereb Blood Flow Metab 16:817–826.

Koster-Hale J, Saxe R (2013): Theory of mind: A neural prediction problem. Neuron 79:836–848.

Kovács ÁM, Téglás E, Endress AD (2010): The social sense: Susceptibility to others' beliefs in human infants and adults. Science 330:1830–1834.

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.

Lloyd-Fox S, Blasi A, Elwell CE, Charman T, Murphy D, Johnson MH (2013): Reduced neural sensitivity to social stimuli in infants at risk for autism. Proc R Soc B Biol Sci 280:20123026.

Lombardo MV, Chakrabarti B, Bullmore ET, Wheelwright SJ, Sadek SA, Suckling J, Baron-Cohen S (2010): Shared neural circuits for mentalizing about the self and others. J Cognit Neurosci 22:1623–1635.

Low J (2010): Preschoolers' implicit and explicit false-belief understanding: Relations with complex syntactical mastery. Child Dev 81:597–615.

Luo Y, Baillargeon R (2007): Do 12.5-month-old infants consider what objects others can see when interpreting their actions?. Cognition 105:489–512.

Luo Y, Johnson SC (2009): Recognizing the role of perception in action at 6 months. Developmental Science 12:142–149.

Ma N, Vandekerckhove M, Van Overwalle F, Seurinck R, Fias W (2011): Spontaneous and intentional trait inferences recruit a common mentalizing network to a different degree: spontaneous inferences activate only its core areas. Soc Neurosci 6:123–138.

Maris E, Oostenveld R (2007): Nonparametric statistical testing of EEG- and MEG- data. J Neurosci Meth 164:177–190.

McKinnon MC, Moscovitch M (2007): Domain-general contributions to social reasoning: Theory of mind and deontic reasoning re-explored. Cognition 102:179–218.

Mitchell JP (2009): Inferences about mental states. Philos Trans Roy Soc B Biol Sci 364:1309–1316.

Mitchell JP, Macrae CN, Banaji MR (2006): Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron 50:655–663.

Obrig H, Neufang M, Wenzel R, Kohl M, Steinbrink J, Einhäupl K, Villringer A (2000): Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults. Neuroimage 12:623–639.

Obrig H, Villringer A (2003): Beyond the visible-imaging the human brain with light. J Cereb Blood Flow Metab 23:1–18.

Onishi KH, Baillargeon R (2005): Do 15-month-old infants understand false beliefs? Science 308:255–258.

Piazza M, Izard V, Pinel P, Le Bihan D, Dehaene S (2004): Tuning curves for approximate numerosity in the human intraparietal sulcus. Neuron 44:547–555.

Premack D, Woodruff G (1978): Does the chimpanzee have a theory of mind? Behav Brain Sci 1:515–526.

Robinson S, Goddard L, Dritschel B, Wisley M, Howlin P (2009): Executive functions in children with autism spectrum disorders. Brain Cognit 71:362–368.

Ruxton GD, Neuhäuser M (2010): When should we use one-tailed hypothesis testing? Methods Ecol E 1:114–117.

Saxe R (2010): The right temporo-parietal junction: A specific brain region for thinking about thoughts. In: Leslie A, German T, editors. Handbook of Theory of Mind. Mahwah, NJ: Taylor and Francis Group. pp 1–35.

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.

Schneider D, Bayliss AP, Becker SI, Dux PE (2012): Eye movements reveal sustained implicit processing of other's mental states. J Exp Psychol Gen 141:433–438.

Schneider D, Slaughter VP, Becker SI, Dux PE (2014): Implicit false-belief processing in the human brain. Neuroimage 101: 268–275.

Scholkmann F, Spichtig S, Muehlemann T, Wolf M (2010): How to detect and reduce movement artifacts in near-infrared imaging using moving standard deviation and spline interpolation. Physiol Meas 31:649.

Scott RM, He Z, Baillargeon R, Cummins D (2012): False-belief understanding in 2.5-year-olds: Evidence from two novel verbal spontaneous-response tasks. Dev Sci 15:181–193.

Senju A (2012): Spontaneous theory of mind and its absence in autism spectrum disorders. Neuroscientist 18:108–113.

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.

Sommer M, Döhnel 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.

Strangman G, Boas DA, Sutton JP (2002): Non-invasive neuroimaging using near-infrared light. Biol Psychiatry 52:679–693.

Strangman G, Culver JP, Thompson JH, Boas DA (2002): A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation. Neuroimage 17: 719–731.

Strangman G, Franceschini MA, Boas DA (2003): Factors affecting the accuracy of near-infrared spectroscopy concentration calculations for focal changes in oxygenation parameters. Neuroimage 18:865–879.

Surian L, Caldi S, Sperber D (2007): Attribution of beliefs by 13-month-old infants. Psychol Sci 18:580–586.

Van Overwalle F, Vandekerckhove M (2013): Implicit and explicit social mentalizing: Dual processes driven by a shared neural network. Front Human Neurosci 7:1–6.

Wellman HM, Cross D, Watson J (2001): Meta-analysis of theory-of-mind development: The truth about false belief. Child Dev 72:655–684.

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.

Yucel MA, Selb J, Boas DA, Cash SS, Cooper RJ (2014): Reducing motion artifacts for long-term clinical NIRS monitoring using collodion-fixed prism-based optical fibers. Neuroimage 85: 192–201.