

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Psychology

2021

Modeling Awareness for the Self and Others

Bio, Branden Joseph

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp017m01bp82b>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

MODELING AWARENESS FOR THE SELF AND OTHERS

BRANDEN JOSEPH BIO

A DISSERTATION PRESENTED TO THE FACULTY OF
PRINCETON UNIVERSITY IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF
PSYCHOLOGY

ADVISOR: MICHAEL S. A. GRAZIANO

SEPTEMBER 2021

© Copyright by Branden Joseph Bio, 2021. All rights reserved.

Abstract

Understanding the minds of others is a crucial ability for humankind. However, one fundamental aspect of this, how the awareness of others is modeled and where this ability is rooted in the brain, is still not well understood. The experiments described in this dissertation investigate how individuals understand and reconstruct the awareness and attention of others and the ways these processes relate to the construction of their own awareness. The study presented in chapter 2 offers evidence that the ways in which individuals use and combine information to assess others' awareness is closely tied to their social cognitive ability. Chapters 3 and 4 provide evidence that an individual's attentional biases seem to transfer when modeling the attention of others, demonstrating a tight relationship between understanding others' attention and one's own. Finally, the studies presented in chapters 5 and 6 suggest that areas of the brain responsible for cognition about others represent rich content of another agent's awareness as well as details about the agent's attention. These studies provide support for the concept that models of the self and models of others have shared mechanisms that may rely on the same underlying neural substrates.

Table of Contents

Abstract	iii
Table of Contents	iv
Acknowledgments	v
Chapter 1. Introduction	1
Chapter 2. Integrative mechanisms of social attention	11
Chapter 3. Attentional biases transfer to judgments of others' awareness	29
Chapter 4. Theory of mind deficits in patients with spatial neglect	48
Chapter 5. Encoding the mental experience of others versus the truth-value of beliefs	58
Chapter 6. Modeling types of attention in the self and others	84
Chapter 7. Conclusion	107
References	110

Acknowledgments

I would like to thank the following individuals for their support throughout my time in graduate school. My adviser, Michael Graziano, whose intellectual curiosity, and passion helped drive me forward. My Reading Committee, Susan Fiske and Diana Tamir, whose questions, and comments lead to the best version of this dissertation. My Oral Committee, Tania Lombrozo and Casey Lew-Williams, who have been great sources of encouragement, mentorship, and intellectual stimulation long before being on any of my committees. My collaborators, especially Caroline Schnakers, without whom Chapter 4 would not have been possible.

I would also like to acknowledge the friends, some of whom I made while at Princeton, that helped me survive this long journey. Karina, thank you for being the glue of the cohort and for giving me a family in graduate school. Andrew, my close friend, officemate, and academic brother, I could not imagine going through this without you and I am so glad to have shared this experience (and an office) with you. Britney, thank you for sticking with me all these years and for always believing in me and cheering me on. Finally, Elyssa, I cannot express in words the extent of my appreciation for all that you have given me but know that you have been a beacon of light and that every day I have spent with you has been better than the one before.

Chapter 1 - Introduction

“For everything begins with consciousness and nothing is worth anything except through it.”

-Albert Camus, *The Myth of Sisyphus*

For thousands of years, investigating the mind belonged exclusively to the realm of philosophy. Armed with well-constructed argumentation, introspection, and thought experiments, philosophers devised many potential explanations for what the mind might be, and how it could come about. Although evidence exists of spiritual beliefs and something akin to personhood in early peoples, the modern concept of a person as a single unified internal subject of their own thoughts and behavior didn't coalesce until Rene Descartes. In his *Principles of Philosophy* (1644/1985) Descartes wrote: “By the word ‘thought’ I understand all that of which we are conscious as operating in us.” It took another 200 years for the question of the mind to become the focus of empirical investigation. As a science of psychology was blossoming in the 19th century, the aspects of the mind suitable for empirical study were also being determined. At that time, consciousness – subjective experience or the daunting *what-it-is-like*, was left as something unknowable, something too big to tackle. It wasn't until the cognitive revolution of the 1950's and the refutation of behaviorism that the door even opened for the possibility of studying consciousness as a phenomenon again. Advancements in computing at this time brought to mind the potential for similarities between the human mind and the computing machine. It would still be several decades before consciousness would be taken up seriously as a valid research topic, but in the 1980's and 90's psychologists and neuroscientists moved from the

modular mind of the cognitive revolution to dare to explore how all of those components could come together to form a cohesive subjective experience.

Since the Global Workspace Theory was first proposed by Bernard Baars (1988), a number of scientific theories have emerged and have attempted to explain what consciousness is, and how the brain creates it (see Block, 2009 for a review). The work contained in this dissertation is directly motivated by one theory of consciousness in specific, the Attention Schema Theory (AST). The foundational ideas behind this theory were first put forth by Graziano and Kastner (2011) as “a construct of the social perceptual machinery.” In brief the theory claims that: (1) awareness is a functionally useful schematic perceptual reconstruction of attention and (2) that the mechanisms responsible for computing information about other people’s awareness are the same ones that compute information about our own awareness. The collection of experiments contained herein test this second, social hypothesis; that the way we construct the awareness of others, how we understand their minds, is intimately tied to how we construct our own mind. Chris Frith put it well in a 2002 paper, pointing out that “A common format for representing mental states that applies to our own as well as others opens up the possibility for communication.” While being able to take the perspective of others and getting a first-hand understanding of their experience allows individuals to communicate effectively, it does so much more than that; it also provides the ability to interact with, predict, and truly understand others’ minds.

Social cognition has long been concerned with how individuals reconstruct the contents of other people’s minds. In this way, the mind, or more specifically, awareness, can be thought of as something akin to a container. Whichever items attention takes hold of, be they perceptions, beliefs, thoughts, memories, etc. become the contents of this container. Social cognition then is

mainly the study of how we reconstruct the contents inside another person's awareness. But there is a more fundamental question buried here, namely, how do people reconstruct the presence of this container, the awareness itself, inside another person? How do we model someone else's attention? Information about someone else's attention is useful in predicting their behavior, whether they are aware of X or Y, and how they might interact with these items. Applying this sort of social perception to oneself provides not only a description of one's own beliefs and thoughts, but also a description of one's awareness of items in the outside environment. It is for this reason that awareness of any perception, not just merely self-awareness, can be understood as a social computation being made for oneself. The five experiments presented in this dissertation get at this question of how people model the attention container inside other people and how that relates to the modeling of their own attention container. Each experiment approaches this question in a different way, but collectively they show that the brain generates a model of other people's attention and awareness through similar mechanisms to the model it generates for the self.

Many successful approaches have been taken to study individuals' relationship to others' minds, each focusing on a slightly different area of cognition. Work in theory of mind has explored how and when humans gain the ability to understand the thoughts of others. Social attention has focused more narrowly on how people track the attention of others in an environment. My work has been inspired by these areas of research. I have adapted methods from theory of mind and social attention to examine how individuals construct the attention and awareness of others and compared how that process relates to the construction of their own attention and awareness. In the next two sections I will briefly review these two areas of

research, not exhaustively, but focusing only on the literature of direct relevance to the experiments in this dissertation.

Theory of Mind

A foundational component of social interactions generally and social cognition more specifically is theory of mind (ToM). First coined by Premack and Woodruff (1978) in their work on chimpanzees, they defined ToM as the ability to “impute mental states to oneself and to others.”

Wimmer and Perner were pioneers in studying how children think about the beliefs of others. In their work, *Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception* (1983), Wimmer and Perner used a cartoon vignette paradigm to test children from the ages of 3-6 years old. This “Sally-Anne task” as it would come to be called would become the gold standard method of testing theory of mind about false beliefs. In addition, they found that the age of 4 years old was roughly when children were first able to start understanding that other individuals could have beliefs different from their own or from the true state of affairs, i.e., beliefs that were false given the situation. The age at which humans develop the ability for ToM is still hotly debated, with some work suggesting it may come online before the age of 2 years old (Buttelmann et al., 2009; Onishi & Baillargeon, 2005; Southgate et al., 2007; Surian et al., 2007).

As ToM was being investigated through the course of healthy development it was also being explored in parallel in children suffering from autism and other developmental disorders. Researchers such as Uta Frith, Alan Leslie and Simon Baron-Cohen explored how children with developmental disorders might reason about others' beliefs. An early paper by Baron-Cohen

(1985) used Wimmer and Perner's paradigm to show that, compared to neurotypical children who were much younger, as well as children with Down's syndrome, children diagnosed with autism were not able to understand that others had false beliefs. Other work came to the same conclusion that individuals with autism were not as capable at determining the mental states of others as their neurotypical counterparts (Leekam & Perner, 1991; Perner et al., 1989; Swettenham, 1990).

Along with the false-belief vignettes previously mentioned, which are used primarily with children, a number of other methods are prevalent in the ToM literature, especially in adult populations. There is a notable split in the literature between verbal and non-verbal methods when measuring ToM capabilities. Written stories, which tend to be popular in neuroimaging studies, often consist of scenarios regarding individuals with false beliefs (Saxe & Kanwisher, 2003; Saxe et al., 2006), those who have incomplete information, or those who need to decipher the intentions or implied meanings of another individual such as in the Hinting Task (Corcoran et al., 1995). Non-verbal assessment tools are also common such as the Reading the Mind in the Eyes Task (Baron-Cohen, 2001), the Awareness of Social Inference Test (McDonald et al., 2003), false-photograph task (Zaitchik, 1990), as well as variations of false belief tasks in cartoon form (Sommer et al., 2007), and inferring mental states from animated shapes (Malhi et al., 2008).

Over the past 25 years ToM has been extensively studied in the human brain using fMRI. Although not the first to explore ToM in the brain, Saxe & Kanwisher's 2003 paper was pivotal and may be the most cited work on the subject. In addition to a few others that preceded it, this work established a consistent set of regions that were activated during ToM reasoning, namely: the right and left temporoparietal junction (TPJ), right anterior superior temporal sulcus (STS),

the precuneus (PC), and the medial prefrontal cortex (MPFC; Castelli et al., 2000; Fletcher et al., 1995; Gallagher et al., 2000; German et al., 2004; Goel et al., 1995; Saxe & Kanwisher, 2003; Vogeley et al., 2001; Xiao et al., 2019). This group of regions has since been deemed the theory of mind network. Since establishing the network involved, research has moved toward exploring the specificity of these regions and differentiating their involvement when thinking about others. Two regions that stick out as the most reliable to activate across all ToM tasks are the TPJ and mPFC. The TPJ most commonly activates in scenarios involving attribution of temporary or hypothetical states to others (Allison et al., 2000; Decety & Lamm, 2007; Saxe & Wexler, 2005; Van Overwalle, 2009) while mPFC seems most strongly activated when thinking of others' mental states compared to one's own (Amodio & Frith, 2006; Baetens et al., 2014; D'Argembeau et al., 2007; Frith, 2007; Northoff et al., 2006; Saxe, 2006).

This non-exhaustive list provides ample evidence that there is an abundant and active literature exploring theory of mind in several different populations ranging from young children to adults, including both neurotypical and neurodivergent individuals, as well as using an extensive array of methodologies from behavioral approaches to fMRI.

Social Attention

I will now discuss some of the background of a related topic, social attention. When discussing a topic as broadly construed as social attention it is important to be clear about what exactly is meant. In the literature, the term 'social attention' is often used in a number of ways. It is sometimes used generally as an umbrella term for attention to other people or social cues, and often conflated with another related concept, joint attention. Joint attention is the coordinated orienting of different individuals toward the same object (Emery et al., 1997; Kristen et al., 2011; Moore & Dunham, 1995; Mundy & Newell, 2007). Although this process is an important part of

social attention, it is not the only consideration of interest that contributes to how people track the attention of others. Several components come together to inform how individuals make sense of others' attention, including eye gaze direction (Friesen & Kingstone, 1998), gestures (Frischen et al., 2007; Langton & Bruce, 2000), head orientation (Langton & Bruce, 1999), as well as other factors such as context to help create the whole picture of what another person has in mind.

Social attention emerges early on in human development (Butterworth & Jarrett, 1991; Carpenter et al., 1998; Farroni et al., 2000, 2004; Hood et al., 1998; Scaife & Bruner, 1975; Young-Browne, 1977), with some studies suggesting infants show sensitivity to eye gaze as early as 5 months (Symons et al., 1998). Head turning to mirror parental direction has been shown in 6-month-old children (Morales et al., 1998). Although not the first to note the separate acts, Mundy and colleagues (2007) delineated a distinction between two types of social attention important in healthy infant development; responding to joint attention, which “refers to infants’ ability to follow the direction of the gaze and gestures of others in order to share a common point of reference” and initiating joint attention, which “involves infants’ use of gestures and eye contact to direct others’ attention to objects, to events, and to themselves.” These two separate processes are important milestones in development. It is clear that social attention is an early feature of human development and one that is crucial for healthy maturation.

Some of the earliest interest in social attention came from a desire to understand deficits in autism and similar developmental disorders that affect social functioning (Baron-Cohen, 1989; Mundy & Sigman, 1989; Pellicano, 2010; Lord & Jones, 2012). The previously mentioned responding to joint attention and initiating joint attention are valuable constructs in their own right but have also become central to the diagnosis of autism (Charman, 2004; Gotham et al., 2007). Impairments specifically in the initializing of joint attention are typically more severe and

are therefore more helpful in diagnosis (Gotham et al., 2007; Kasari et al., 2008; Mundy et al., 1994). The extent of an individual's social attention ability has also been found to be larger in individuals with higher social competence as measured on the autism quotient, a measure of the expression of autistic traits in adults (Bayliss & Tipper, 2005).

In another approach to social attention, Friesen and Kingstone (1998) modified a cueing paradigm from Posner (1980) which used an arrow as an attentional cue, followed by a response target at either the cued location or at an uncued location. Friesen and Kingstone turned this paradigm into a test of social attention by replacing the central arrow with a face gazing one way or the other. This task allows for a measure of the cue's ability to direct attention; when a target at a congruent location is detected faster, or more accurately, than a target at a non-congruent location, the cue is better able to shift attention. Many others have since used this task as well (Driver et al., 1999; Friesen et al., 2004; Frischen et al., 2007; Hayward & Ristic, 2013; Hood et al., 1998; Tipples, 2008).

Critics of experiments performed in a controlled lab setting have taken to studying social attention in more ecologically valid ways, in more naturalistic settings. Laidlaw and colleagues (2011) had subjects sit in a waiting room with either an in-person confederate or a video feed confederate and measured eye movements with a mobile eye tracker. In a pair of studies, Gallup and colleagues measured eye gaze of pedestrians directed at a camera on the street (Gallup et al., 2012a, 2012b). The combination of work in-lab as well as more naturalistic studies have shown that social cues engage and shift attention, and that in day-to-day life, social attention is modulated by the presence of others, the potential for social interaction, and the relevance of social information.

Similar to ToM research, social attention has been widely investigated using fMRI. Unsurprisingly, there is a fair amount of overlap in the areas involved in thinking about others' mental states and in tracking other people's attention. Eye gaze, the key component of joint attention, has repeatedly been tied to activation in STS (Calder et al., 2007; Friesen & Kingstone, 2003; Kingstone et al., 2004; Pelphrey et al., 2004). When compared to the same Posner cueing paradigm using arrows, eye gaze differentially activated STS (Friesen & Kingstone, 2003; Kingstone et al., 2004). Another commonly found region involved in both ToM and social attention is mPFC (Bristow et al., 2007; Calder et al., 2002; Williams et al., 2005). Two components of the dorsal attention network implicated in goal-directed and exogenous shifts in attention, the superior parietal lobule (SPL) and the frontal eye fields (FEF), are also found to contribute to tracking eye gaze. Indeed, regions involved in social attention seem to be a combination of both the ToM network as well as areas active when engaging one's own attention.

Models of the Self and Models of Others

Theory of mind and social attention are both processes that consist of modeling the minds of others. They both involve thinking about thinking, that is, having thoughts themselves as their content, which puts them both in the category of metacognition (Bartsch & Estes, 1996; Chiu & Kuo, 2009; Decety & Lamm, 2007; Ebert, 2015; Kuhn, 2000; Petty et al., 2007; Proust, 2007; Sodian & Frith, 2008). Modeling the mental states of others provides the ability to predict the future behaviors and actions of those around you, which is crucial to successfully engage in a social world.

Vogeley and colleagues were early in considering the close relationship between modeling others and modeling the self. In their 2001 paper (Vogeley et al., 2001), ToM was

compared with what the authors referred to as ‘self-perspective’. In both behavioral and fMRI studies, participants were asked to report on mental states from the perspective of themselves and others. It was found that there were both areas of overlap (prefrontal cortex) and distinct regions of activity for the self (rTPJ) and other conditions (rACC). Saxe and colleagues (2006) took a more cognitively focused approach and looked at ToM with respect to self-reflection. Using ToM stories and self-directed trait words, areas of overlap (prefrontal cortex and precuneus) and distinct areas for self (medial PFC) and other (temporo-parietal areas) activations were found. A meta-analysis of fMRI studies of self-recognition and theory of mind by van Veluw & Chance (2014) also found overlap of self and other processing as well as distinct areas for each. There is an ample literature comparing the self and others in many domains, including but not limited to emotion (Ochsner et al., 2004; Thornton & Tamir, 2017), perception (Samson et al., 2010; Vogeley et al., 2004), agency (Farrer & Frith, 2002), and decision making (Tamir & Mitchell, 2013).

The studies reviewed in this introductory chapter paint a clear picture that researchers have long suspected a connection between the way humans model their own minds and the minds of others. The goal of this dissertation is to add to this rich existing literature by shedding light on the shared mechanisms of modeling attention and awareness for the self and others and to provide evidence that they are an overlooked but crucial part of social cognition.

Chapter 2 - Integrative mechanisms of social attention

Introduction

Eye gaze has long been used as the primary measure of social attention (Driver et al., 1999; Freeth et al., 2013; Friesen et al. 2004; Frischen et al., 2007; Langton, 2000). Researchers have used eye gaze because of its powerful ability to shift other agents' attention (Bayliss et al., 2010), the speed and automaticity at which it is processed (Langton et al., 2000; Underwood & Foulsham, 2006), as well as the evidence that there are specific neural systems biased toward processing eye gaze (Birmingham & Kingstone, 2009; Nummenmaa & Calder, 2009). Although this approach has shown to be useful, it is very limited and simplifies the act of understanding someone's attention, the complex process of modeling someone's deployment and use of attention, to the mere act of tracking their gaze.

A visual cue commonly used in other areas of social cognition is facial expression due to the ubiquity with which it occurs and the sheer amount of information that can be conveyed through this behavior (Adams et al., 2017; Carroll & Russell, 1996; Ekman, 1993; Hietanen & Leppanen, 2003; Hori et al., 2005; Nook et al., 2015; Vuilleumier, 2002). But facial expression can provide information about more than just emotional state alone; it can also give clues to what is contained in someone's mind. If a person turns toward you with a horrified look on their face and you can see over their shoulder a building is on fire, you might have strong reason to believe that person is aware of that occurrence.

It is important to consider that in the world, visual cues don't exist in a vacuum. You hardly ever see a person looking intently at something with a neutral face or see someone with an intense expression without looking at anything in particular. These processes co-occur and as

such may be used in concert to inform individuals about the attentional state of others. Therefore, it is worthwhile not just to explore how each of these cues may be useful for tracking another person's attention, but also to consider the potential integration of cues (Pecchinenda et al., 2008), which Jamil Zaki suggested was necessary to “formally model the ways that we perceive others' minds based on complex social information” (2013). To date, attention has typically been studied using eye gaze alone; this chapter describes a study on how individuals use eye gaze in combination with facial expression to model the attention and awareness of others.

Previous studies have looked at interactions between gaze and facial expression before but have done so in a different manner than the current approach (Hietanen & Leppänen, 2003; Hori et al., 2005). In those studies, expression was not set up to provide specific information about the attentional state of the agent involved. In the present study, the expression of the face could be congruent with or incongruent with an object, just as the gaze could be toward or away from the object. In this way, facial expression and gaze were potentially equally informative about whether the face was attending to an object. Previous studies on the interaction between gaze and attention generally used stimuli in which gaze was the primary cue to attention and facial expression was a secondary, modulating factor (Hietanen & Leppänen, 2003; Hori et al., 2005). We suggest that in real-life situations, though gaze may indeed often be the primary cue to someone else's attention, sometimes other cues are equally or even more informative.

To help fill the gap in the literature and explore how people interact with different available cues to judge someone else's attentional state, we employed a paradigm from a prior study from our lab. This paradigm was chosen because it restricts the social cues available in a visual scene to those of interest, while allowing each variable to provide equal amounts of information about the target agent. The attribution-of-awareness task used by Kelly and

colleagues asked subjects to view a (cartoon) person on-screen who was either looking toward an object or away from it and had either a happy expression or an agitated one. The objects shown were either positive in nature or negative, and therefore could either match the expression the face had (happy+positive / agitated+negative) or mismatch it (happy+negative / agitated+positive). Subjects were then asked to rate how aware the person on-screen was of the object. This setup gave subjects two visual cues that they could use either individually or in combination to inform their judgments about others' awareness. Subjects could use the direction of the cartoon individual's eye gaze as the main decider of where their attention was directed, they could use the congruency of the facial expression, or some combination of both factors.

Although the standard view of social attention gives the bulk of the theoretical weight to eye gaze (Driver et al., 1999; Freeth et al., 2013; Friesen et al. 2004; Frischen et al., 2007; Langton, 2000), we believe that individuals compute a more complex model of others' attention than merely following where another individual is looking. We propose that an individual's ability to model the awareness and attention of others is a crucial part of their social cognition, and therefore individuals who are more adept at reading and engaging in social situations should be more likely to integrate multiple social cues together when making assessments of another person's awareness.

To measure individuals' social cognitive ability, we tested subjects using two common assessments. The Reading the Mind in the Eyes Test (RMET; Baron-Cohen et al., 1997), which is widely used as a perceptual assessment for emotion recognition; and the Hinting Task (HT; Corcoran et al., 1995) which is a cognitive theory of mind assessment, were chosen to give us two independent measures of social cognition that assess understanding the minds of others using very different methods. The RMET consists of a photograph of the eyes of an individual,

which participants are asked to match to the emotion that individual is expressing. The HT consists of short stories, where a scenario is presented in which two individuals' interaction leads to a situation where one person says or does something that hints some desire or intention to the other person, which the subject must then identify.

If another person's eye gaze or expression is the primary factor that contributes to understanding how others' attention is employed, individuals who employ this strategy should perform best on the RMET and HT. If people monitor each other's attention by constructing a deeper model, one that combines multiple cues as we propose, it is expected that social cognition scores for the two tasks should correlate with the ability to integrate cues about other people's focus of attention.

Methods

Participants

All participants provided informed consent and all procedures were approved by the Princeton Institutional Review Board. Eight hundred and thirty healthy human volunteers were recruited through Amazon Mechanical Turk (229 females, aged 18-70, mean age = 37 years, SD = 10) and were paid \$0.10 per minute of the study. Two hundred and thirty-eight participants were excluded prior to analysis for failing to meet the threshold of answering 80% of trials on the attribution-of-awareness task, or for providing incoherent answers on the HT which suggested the presence of an automated, non-human responder. Therefore, 582 subjects were included in analyses.

Stimuli Validation

Prior to testing on the attribution-of-awareness task, stimuli were validated in an independent sample of 56 individuals recruited through Amazon Mechanical Turk. In the stimuli validation, volunteers rated the appropriateness of the expression of a face looking at a set of 96 objects. On each trial, an object was shown with a cartoon head facing it. Objects consisted of natural items such as foods and animals as well as artificial items and events. The cartoon either had a happy expression or an agitated expression when looking at the object. Participants were then asked to choose on a 3-point scale whether the face's expression was: "inappropriate", "neither inappropriate nor appropriate" or "appropriate" for the object that it was paired with. The 30 most unambiguous positive objects (rated most highly appropriate with a happy expression and most inappropriate with an agitated expression) and 30 most unambiguous negative objects (rated most highly appropriate with an agitated expression and most inappropriate with a happy expression) were included in the final task. Objects were split into 'positive' and 'negative' categories in order to create congruent and incongruent conditions for the attribution-of-awareness task described below.

Attribution-of-Awareness Task

The 'attribution-of-awareness' task included in this study was adapted from Kelly et al. (2014). Participants were asked to judge the extent to which an agent on-screen was aware of an object (see figure 1). It should be noted that whereas the figure shows the relative sizes and positions of images in the display, the absolute size in degrees of visual angle could not be controlled because of the online testing procedures. The task consisted of 240 trials over 8 blocks in which a cartoon head was displayed on-screen simultaneously with a picture of an object. Participants were asked to 'respond whether you think the cartoon face looks "not aware",

"somewhat aware" or "very aware" of the object'. Judgments were made on a 3-point scale from 'Not Aware' coded as -1, 'Somewhat Aware' coded as 0, and 'Very Aware' coded as 1.

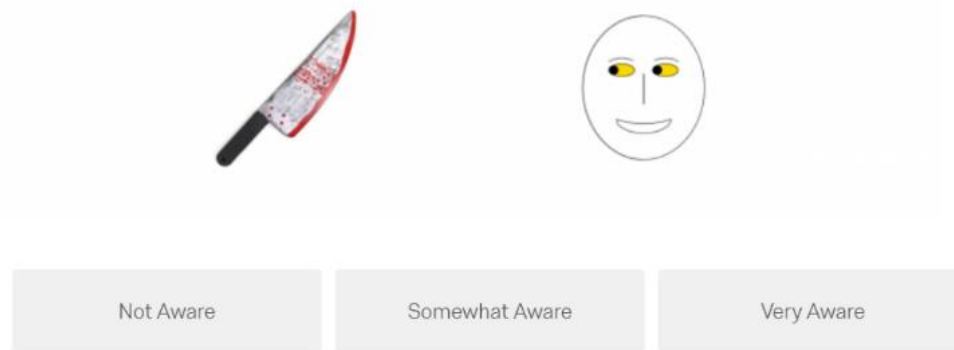


Figure 1. Paradigm for the attribution-of-awareness task. Of the four possible stimulus conditions, shown here is an example of the gaze toward, expression incongruent trial (gaze+expr-). The other conditions were gaze toward, expression congruent (gaze+expr+); gaze away, expression incongruent (gaze-expr-); gaze away, expression congruent (gaze-expr+).

For half of the trials the cartoon's eyes were directed toward the object and for the other half of trials the eyes were directed away from the object. The expression of the face was happy for half of trials and agitated for half of trials. The facial expression was either congruent or incongruent with the valence of the object. This 2X2X2 design was collapsed into four conditions of interest for the purpose of analysis: gaze toward, congruent expression (gaze+expr+); gaze toward, incongruent expression (gaze+expr-); gaze away, congruent expression (gaze-expr+); and gaze away, incongruent expression (gaze-expr-). Gaze direction and agent expression were the sole cues available to inform participants' judgments on whether the cartoon agent was aware of an object.

Social Cognition Tasks

General aptitude for social cognition was measured using two common assessment tools. We included two tasks that have been widely validated as reliable tasks for assessing theory of mind; the Reading the Eyes in the Mind Task (Baron-Cohen et al., 1997; 2001) and the Hinting Task (Corcoran et al., 1995; Greig et al., 2004). The RMET consists of 36 photographs of faces restricted to the eye area which subjects label with an emotion from a small set given on each trial. This task measures a subject's ability to recognize the emotional state of others from their eyes alone, assessing how well individuals can 'read' the mental state of others given limited social cues. For each correct matching of the emotional state with the expression a subject is given 1 point up to a maximum of 36.

The Hinting Task (HT) used here was an adapted version for American participants (altered from the original British version to include more familiar vocabulary and examples). This task consists of 10 short written scenarios where two individuals interact, and one implies something to the other. For example: "Scott and Eric are coworkers who enter the elevator at the same time. Scott's arms are filled with paperwork. He says to Eric, 'Sixth floor please, I'm late for a meeting.'" Subjects are then asked a question, specific to each trial, that probes what was implied by the characters, such as: "What does Scott really mean when he says this?" Each question earns 2 points for a correct answer for a maximum score of 20 points. In previous versions, subjects were given a second chance to respond to incorrect responses to earn 1 point, but given the online format, here we chose to limit subjects to one try at each question, thus making scoring 2 points for a correct response and 0 for an incorrect response.

Results

K-Means Clustering of the Attribution-of-Awareness Task

In the attribution-of-awareness task, each subject was characterized by four means, or the mean response to each of the four task conditions. For example, a subject might generally score the gaze-expr- category with a low awareness near -1, the gaze+expr+ category with a high awareness near +1, and so on. Different subjects showed different patterns across the four means. Despite the fact that only two visual cues were available to guide subjects' decision making on the attribution task, a wide array of response patterns was found. To better understand the responses across subjects, we employed k-means clustering. K-means clustering is an unsupervised learning algorithm that used each individual subject's four means to sort participants into clusters based on similarity in their response patterns. For the dataset collected, five unique response groups were found as shown in Figure 2.

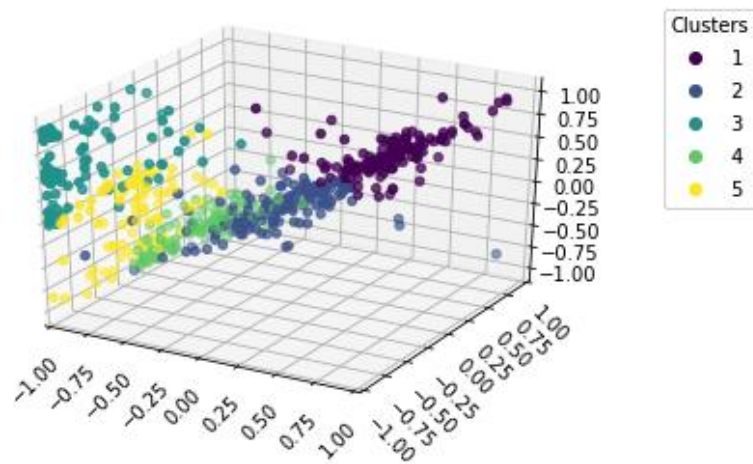


Figure 2. K-means clustering results for the attribution-of-awareness task. Each data point represents a single subject. While only three axes are visually displayed here, the data are actually distributed in four dimensions, corresponding to the four means obtained for each subject. The five colors represent the five k-means clusters.

The mean response pattern for each of the five clusters is shown in Figure 3. In the first cluster, subjects showed a similar, elevated mean across all four trial conditions. These subjects evidently judged every face to be a little aware of the object, without considering either the gaze or expression cue. This cluster contained 132 subjects (22.4% of total).

The second cluster of subjects was similar to the first, in that they were insensitive to either cue, providing similar ratings across all conditions. However, this second cluster consistently rated the faces as only moderately aware. This cluster contained 149 subjects (25.3% of total).

The third cluster of subjects relied primarily on the gaze of the agent (whether gaze was aimed toward or away from the object) to inform their judgments. When the gaze was away from the object, these subjects rated the cartoon as mostly unaware of it; and when the gaze was toward the object, the subjects rated the cartoon as mostly aware of it. This cluster contained 101 subjects (17.2% of total).

The fourth cluster of subjects relied primarily on the expression of the agent (whether the expression was congruent or incongruent with the valence of the object). When the expression mismatched the valence of the object, these subjects rated the cartoon as mostly unaware of it; and when the expression matched the valence of the object, the subjects rated the cartoon as mostly aware of it. This cluster contained 115 subjects (19.6% of total).

Finally, the fifth cluster of subjects integrated both the gaze and expression cues to inform their judgments. They rated the cartoon as mostly unaware of the objects only when both gaze and expression cues mismatched the object; they rated the cartoon as mostly aware of the object only when both gaze and expression cues matched the object; and when the two cues

conflicted, the subjects appeared to balance the weight of each cue, resulting in an intermediate rating of awareness. This cluster contained 91 subjects (15.5% of total).

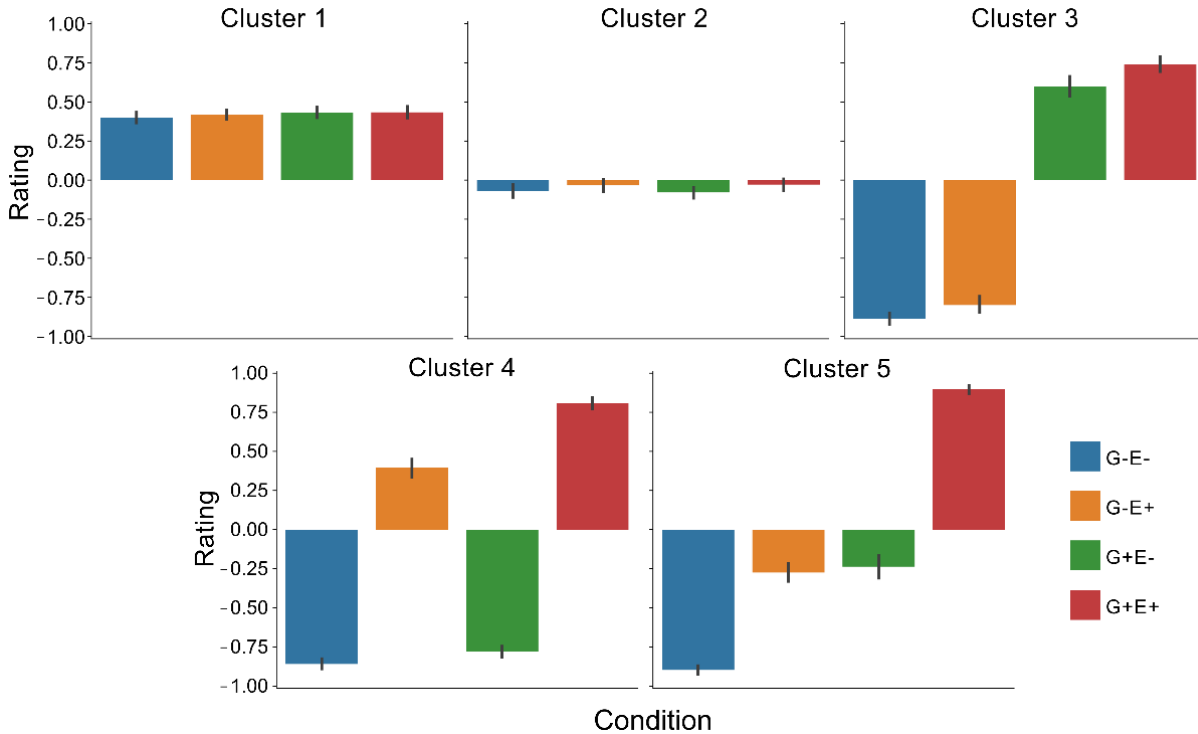


Figure 3. Average response pattern of each cluster output from k-means clustering.

RMET and Hinting Task Performance

The average score on the RMET was 21.1, with a standard deviation of 8.36 and range from 5-37. Performance on the RMET for this study was slightly lower than past studies which used this measure (Fertuck et al., 2009 [M = 25.00, SD = 3.63]; Jankowiak-Siuda et al., 2016 [M = 24.98, SD = 4.53]), although the current study had a similar average score to at least one study (Kynast et al., 2020 [M = 23.4, SD = 10.53]).

The average score on the HT was 10.31, with a standard deviation of 8.18 and range from 0-20. The HT scores collected here are difficult to compare to scores from previous studies, due to modifications made to the scoring of this dataset.

Although the RMET and HT are independent measures of social cognition and assess participants in very different ways, focusing on different aspects of thinking about others, scores for each assessment were highly correlated with one another ($r = 0.82$, $p = 2.15e-19$) showing strong consistency between measures.

RMET and HT scores for each k-means cluster

Figure 4 shows the average score on the two social cognition tasks for each of the k-means clusters of subjects. RMET scores are shown in Figure 4A, and HT scores are shown in Figure 4B. Note that the five clusters form two larger groups. In one group (clusters 1 and 2), subjects performed poorly on both social cognition tests (RMET: $M = 14.42$, $SD = 4.95$; HT: $M = 4.21$, $SD = 5.61$). In the second group (clusters 3, 4, and 5), subjects performed better on the RMET and HT (RMET: $M = 27.36$, $SD = 5.61$; HT: $M = 16.27$, $SD = 4.56$). Moreover, the difference between these two groups of subjects was significant (RMET: $z = 29.41$, $p < 0.001$, two-tailed; HT: $z = 22.14$, $p < 0.001$, two-tailed).

Though clusters 1 and 2 both performed poorly on the social cognition tasks, we found a significant difference between them, with cluster 2 performing slightly better than cluster 1 (RMET: $t(4) = 2.26$, $p = 0.02$, two-tailed; HT: $t(4) = 3.27$, $p = 0.0012$, two-tailed).

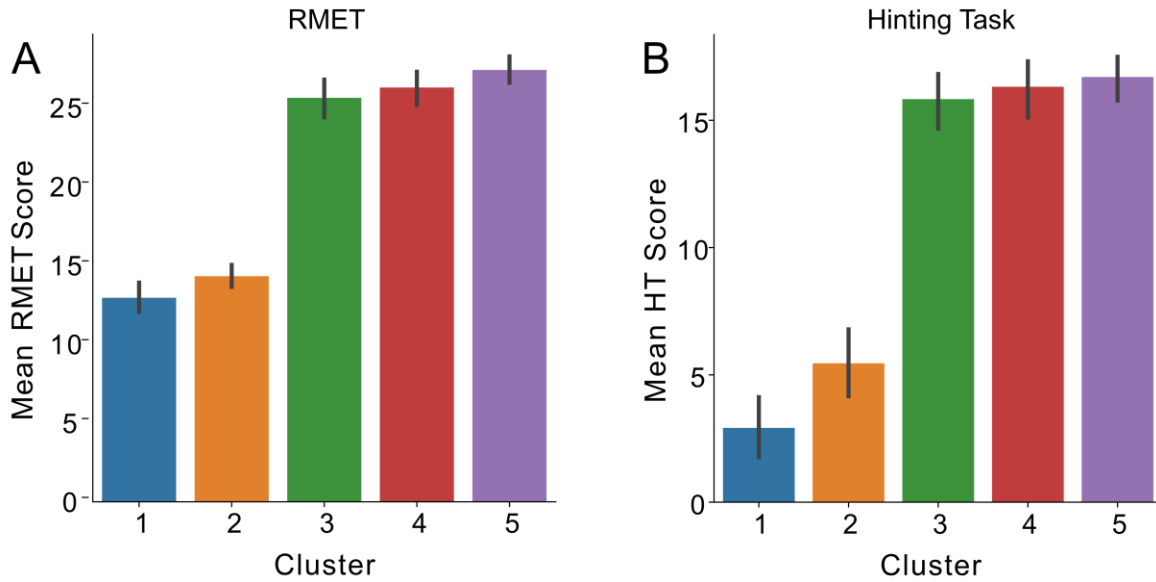


Figure 4. Average (A) RMET score and (B) Hinting Task score for each cluster.

Cluster 1 and 2, as described above, are characterized by an insensitivity to the gaze and expression cues. Subjects in these groups tended to ignore the visual cues when rating the awareness of the cartoon face for the object. It may not be surprising that subjects that ignore social cues available to them also perform poorly on social cognition assessments. However, it cannot be ruled out that these subjects were simply not well engaged in the task and were not paying attention. In an online, remote testing platform, unengaged subjects always pose a risk to experiments. Poor attention would account for why these subjects tended to perform poorly on all three tasks. Although we already eliminated subjects that showed clear evidence of poor engagement with the attribution task, we cannot rule out that many unmotivated subjects remained in the sample. It is of course also possible that these subjects did attend, did try, but simply had very poor social cognition skills. We cannot distinguish between these possibilities, making the interpretation of clusters 1 and 2 difficult.

Do Strategies Show Differences in Social Cognition Performance?

We hypothesized that subjects who integrated the two cues together (cluster 5) should show the best performance on measures of social cognition, whereas subjects who relied on only one or the other cue (clusters 3 and 4) would show poorer social cognition scores. However, the results did not support this hypothesis, as shown in Figure 4. Although cluster 5 showed a marginally higher score in both the RMET and HT over clusters 3 and 4, this difference was not statistically significant. We wondered if perhaps the reason for this negative finding was that the k-means clustering was imprecise, mixing different response strategies to some degree. We therefore devised a new analysis to quantify how much each subject relied on a gaze only strategy, an expression only strategy, and an integration-of-cues strategy.

Figure 5 shows three hypothetical, maximal response strategies. Consider Figure 5A first. Here the strategy relies on the gaze cue only. When gaze is away from the object, a hypothetical subject would rate the cartoon as totally unaware of the object. When gaze is toward the object, a hypothetical subject would rate the cartoon as totally aware of the object. One way to quantify how much a particular subject relies on a gaze strategy is to compute a difference score between the subject's actual response and the optimal gaze strategy represented in this figure. For each of the four conditions shown in Figure 5A, the subject's own score was subtracted from the maximal score. The absolute value of each of those four numbers was then added together. The result was a score for each subject that varied from 0 to 8, in which 0 indicates that the subject perfectly matched the maximal response pattern of using the gaze cue only, and 8 indicates that the subject was as far as possible from the maximal response pattern of using the gaze cue only.

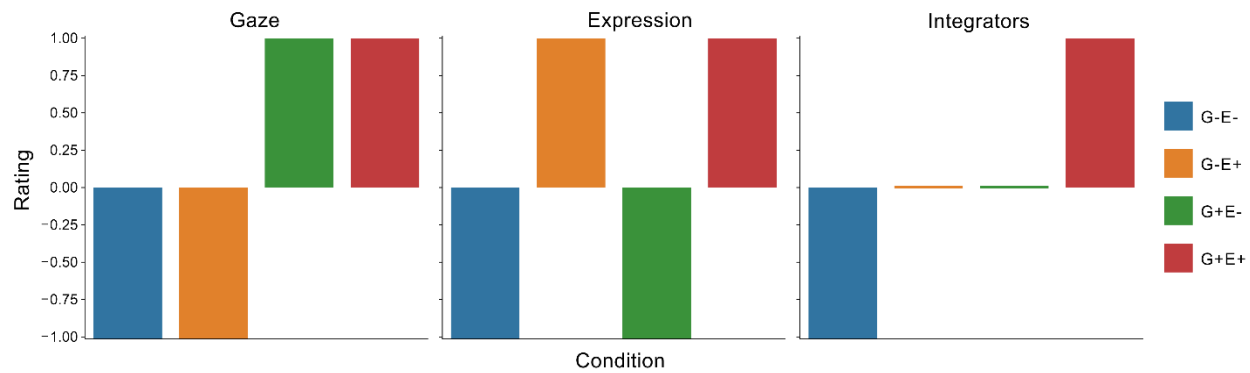


Figure 5. Maximal strategy response patterns for the attribution-of-awareness task.

Similarly, figure 5B shows a maximal response for a hypothetical subject who relies only on facial expression to solve the task. A difference score was computed for each subject to indicate how well the subject matched a response strategy that relied only on facial expression.

Finally, figure 5C shows a maximal response for a hypothetical subject who relies on a perfect integration of the two cues to solve the task. A difference score was computed to indicate how well each subject matched this integrator response strategy.

With these three scores (integrator difference score, gaze difference score, and expression difference score), we could perform correlation analyses, asking whether people who adhered relatively more closely to a particular strategy would also tend to score higher on the social cognition tests. To radically minimize the possibility of inattentive or unmotivated subjects contaminating the results with low scores (as may have occurred in clusters 1 and 2 discussed above), in the following analyses, we split subjects into those who scored above and those who scored below the median, on each of the social cognition tasks. We discarded those who scored below the median. In this way, subjects who showed poor responses were eliminated. We acknowledge that, in this way, we may have discarded many subjects who were fully attentive but simply bad at social cognition. We hoped, however, that this more conservative method of

selecting only the upper 50% of scorers would mostly eliminate the noise of disengaged online subjects and allow for a more straightforward interpretation of the data.

Figure 6A shows that, among the same group of subjects, those who most closely followed a gaze strategy did not perform significantly better on the RMET test of social cognition. The relationship was not significant ($F = 1.52$, $p = 0.22$).

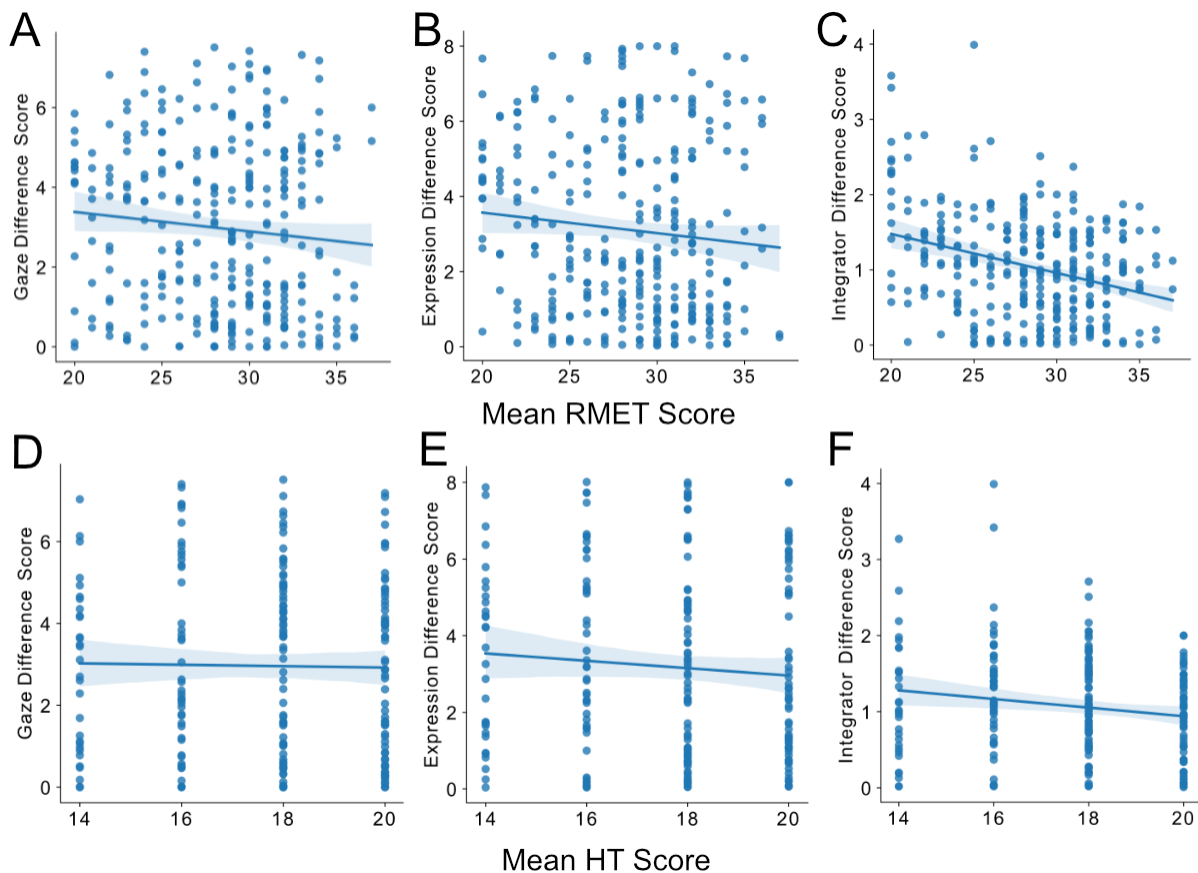


Figure 6. Results for the difference score by social cognition comparison for each of the three strategy types (gaze, expression, integrator). Shown here are the upper half of subjects on a median split of social cognition scores. (A) Strategy difference score correlated with average RMET score for each participant. (B) Strategy difference score correlated with average Hinting Task score for each participant.

Figure 6B shows that, among the same group of subjects, those who most closely followed an expression strategy also did not perform significantly better on the RMET test of social cognition. The relationship was not significant ($F = 0.25$, $p = 0.61$).

Figure 6C shows the correlation between the integrator difference score and the RMET score. Lower numbers on the integrator difference score indicate subjects who relied more on a strategy of integrating cues together. Higher scores on the RMET indicate subjects with better social cognition. The graph shows a highly significant negative correlation. Subjects who most closely followed a cue-integration strategy also performed better on the RMET test of social cognition. The relationship was highly significant ($F = 9.79$, $p = 0.0019$).

Figure 6D, E, and F show the results of a similar analysis, but with respect to the scores on the HT. The results are similar (Gaze: $F = 0.58$, $p = 0.45$; Expression: $F = 1.68$, $p = 0.20$; Integrator: $F = 14.06$, $p = 0.0002$). For both RMET and HT, a relationship can clearly be seen between how closely subjects adhered to a cue-integration strategy and how well subjects performed on an independent test of social cognition. The data therefore suggest that the ability to integrate cues together, to judge the awareness of others, is part of the larger suite of abilities that make up good social cognition.

Discussion

The present study tested participants in three tasks: an attribution-of-awareness task, and two social cognition measures, the RMET and the HT. The attribution-of-awareness task measured how each participant used information to assess whether another agent was aware of a specific object. The RMET is a perceptual assessment of emotion recognition, and the HT is a cognitive theory-of-mind assessment. Although they use very different methods, the RMET and

HT are both measures of social cognition. Performance on each of these social cognition tasks was significantly correlated with the attribution-of-awareness task. The more that subjects relied on a strategy of integrating cues in the attribution-of-awareness task, the better their scores on the RMET and the HT.

The primary finding in this experiment is that when modeling or representing the attentional state of others (in this case, assessing whether a specific object is currently in the mind of another person), people are able to use multiple cues, and are not limited to a simple use of gaze as the sole cue. Modeling the attention of others is not just a matter of tracking gaze, but of using multiple cues to build a deeper model. Moreover, people who show greater tendency to integrate visual cues to arrive at a judgement about another person's attention also score higher on independent social cognition tests, suggesting that this integrator method of modeling attention is part of the larger skill set of social cognition.

These results are consistent with studies of social cognition that explore usage of multiple cues in myriad contexts. Cue integration of perceptual information in social contexts has been found in learning, both at developmental stages (Repacholi et al., 2014) as well as in adults (Parnamets & Olsson, 2020). There is neural evidence of cue integration (Pehrs et al., 2017), and it has even been found in the context of evaluating emotional states (Meeren et al., 2005). Evidence seems to support not just the benefits of using multiple cues in social situations, but specifically the integration of those cues together to effectively navigate and understand the social world and the minds of others. Here, we have found that integrating multiple cues together when modeling the awareness of others is a successful strategy employed by those who tend to be better at social cognition.

In summary, this study found that subjects are not monolithic in the way that they use information to decode the attentional state of others, but rather use a surprising range of strategies. Subjects with higher social cognition, as measured by two assessments, the RMET and the HT, were more likely to integrate multiple sources of information together to inform their judgments. From these results it appears as though individuals who are more adept at social cognition are better at or at least more likely to incorporate more perceptual information from a social environment into their schema of another individual's awareness.

Chapter 3 - Attentional biases transfer to judgments of others' awareness

Introduction

Many people show a small but consistent bias in visual processing, such that judgements along the left-right axis favor one side, which is referred to as pseudoneglect (Brooks et al., 2014; Jewell & McCourt, 2000; Milner et al., 1992; Szczepanski & Kastner, 2013). Typically viewed at the population level, this bias is often subtle and can vary depending on the group being studied, and likewise some individuals can show a clear spatial bias with some people exhibiting a left bias and others exhibiting a right bias. In the present study, we asked whether this bias in one's own spatial processing is mirrored in one's social cognition. If you are biased toward processing your own right side of space, are you also faster at recognizing when someone else processes an object to his or her right side? Graziano and Kastner (2011) have previously suggested that one's own spatial limitations are reflected in one's social perceptions. This chapter describes a behavioral experiment on human subjects, which was published as *Projecting one's own spatial bias onto others during a theory-of-mind task* (Bio et al., 2018).

In the current study, each participant was tested on two tasks. One task was a variant of the line bisection task (Brooks et al., 2014; Jewell & McCourt, 2000; Milner et al., 1992; Szczepanski & Kastner, 2013), used to measure the participant's spatial bias. The second task was a modified version of the theory-of-mind task originally developed by Wimmer and Perner (1983) which was used to measure the participant's social cognition with regards to a cartoon character. The participant had to decide whether the designated cartoon character "thought" that a ball was in box 1 to the left, or box 2 to the right, of the cartoon character. Crucially, the cartoon was rotated such that what was left and right with respect to the cartoon was up and

down with respect to the participant. This rotation therefore prevented the participant's own left-right bias from directly aligning with left and right with respect to the cartoon head. Reaction times were measured in the theory-of-mind task, and we assessed whether participants were faster at responding when the cartoon character supposedly processed an object to its left or its right.

The false-belief task has become a common tool to test social cognition (Baron-Cohen et al., 1985; Turner & Felisberti, 2017; Wellman et al., 2001; Wimmer & Perner, 1983). It typically interleaves two trials types: true-belief and false-belief. False-belief trials of the current task, trials when the participant can see that the ball is in one box but the cartoon character, whose vision is blocked, should think it is in the other box, require social cognition to solve correctly. The participant must consider what the cartoon character believes rather than what is literally true. However, in the true-belief trials, when the participant can see that the ball is in one box and the cartoon character also “believes” the ball is located in the same box, the participant could in principle determine the correct answer by low-level cues (such as the absence of a barrier blocking the cartoon character's sight) and then solve the task without imputing any specific belief to the character. In these trials, the participant could report which box the ball is actually in, thus performing a simple visual task, without ever considering which box the character “thinks” it is in. The true-belief trials therefore might or might not involve some level of social cognition. Only the false-belief trials ensure that the participant engages in social cognition.

For these reasons, in the present experiment, we made a prediction with respect to the false-belief trials. We predicted that people who had a natural bias toward processing their own left side of space would be quicker to respond when the cartoon “thinks” an object lies to the left side of its face, and likewise for a rightward bias. This relationship was predicted specifically for

the false-belief trials in the theory-of-mind task. In the true-belief trials, in which social cognition could not be ensured, we could not make a similar prediction, but the outcome may still provide useful information or insightful interpretation.

Methods

Participants

All participants provided informed consent, and all procedures were approved by the Princeton Institutional Review Board. Seventy-nine participants were tested (55 women, 18-55 years old, median age = 21, normal or corrected-to-normal vision). Each participant performed two tasks: the line bisection task and the theory-of-mind task. The order of tasks was randomized between participants. Criteria for excluding participants, based on task performance, are described below in the sections specific to each task.

Participants sat stabilized by a chinrest 30 cm from a computer screen and used keypresses on a standard keyboard for behavioral responses. Visual stimuli were presented on-screen using Matlab and the Psychophysics Toolbox (Brainard, 1997; Kleiner, Brainard & Pelli, 2007).

Line Bisection Task

In cases of clinical neglect caused by traumatic brain injury or stroke, patients are often tested with the line bisection task (Corbetta, 2014; Halligan et al., 2003; Heilman & Valenstein, 1979). The patient is given a piece of paper with a long horizontal line and asked to draw a vertical transector that bisects the line. A patient with left-sided neglect will typically draw the transector to the right of true center, as though the longer left side appears shortened. In the non-clinical population, a simple line bisection test is typically not sensitive enough to detect

pseudoneglect. A speeded variant of the task with the transector pre-drawn is often used, sometimes called a landmark task (Brooks et al., 2014; Fink et al., 2000; Jewell & McCourt, 2000; Milner et al., 1992; Szczepanski & Kastner, 2013). Figure 7 shows the paradigm used, modified from Szczepanski and Kastner (2013).

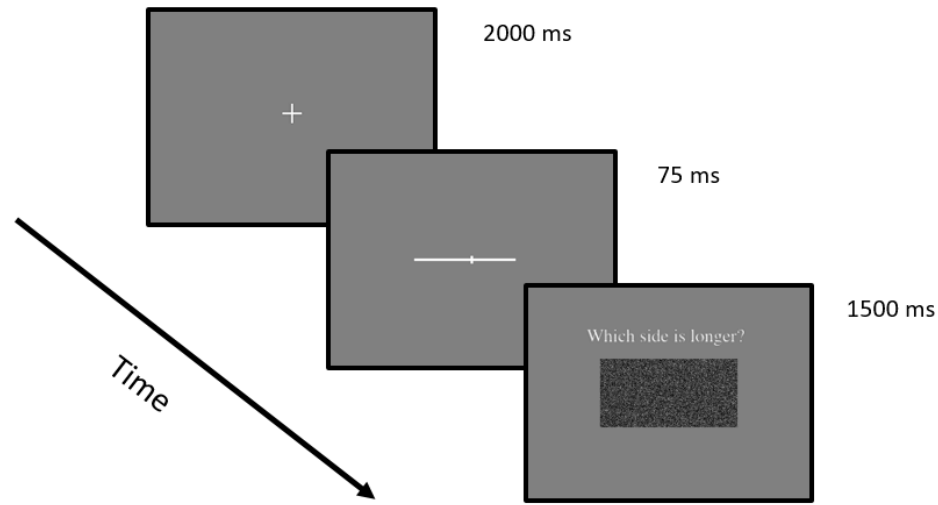


Figure 7. Paradigm for the line bisection task. Participants saw a brief presentation of a transected line and judged which side was longer.

Each trial began with a white fixation cross on a gray background. Participants were instructed to fixate on the cross. After 2000 ms, the cross disappeared and a white horizontal line with a vertical transecting line was displayed for 75 ms. The reason for the brief display was that people are typically extremely sensitive to the left-right asymmetry used in this task. In pilot tests, we settled on a presentation time that reduced performance and resulted in a useful psychophysical curve.

The vertical transecting line was 2 degrees tall in visual angle and was always displayed at the midline of the screen. The horizontal line could be one of four lengths (20, 21, 22, or 23 degrees) and was positioned at one of 11 possible locations (centered or offset to the left or right in 0.2-degree increments, with a maximum offset of 1 degree). This resulted in 44 total stimulus configurations.

After the 75 ms display of the transected horizontal line, a rectangular visual mask (random black and white pixels, 28 degrees in width by 14 degrees in height) covered the area where the stimulus had been. The purpose of the mask was to reduce performance to a range in which a useful psychometric curve could be obtained. Above the mask, a question was presented: “Which side is longer?” Participants were given 1500 ms to respond and were instructed to press one of two keys on a keyboard to indicate whether the horizontal line was longer on the left or right of the vertical transector. Trials where participants exceeded the response window were not included in analyses, however participants responded on 99.6% trials. After the response window, the next trial began immediately.

Participants performed 176 trials broken into 4 blocks of 44 trials each. All trial types were balanced and randomly interleaved. The 4 line lengths and 11 offsets yielded 44 unique trials, however, trials of different line lengths that corresponded on offsets were collapsed together during analysis, yielding 11 analysis conditions. Thus, participants performed 16 trials per condition. The task typically required five minutes to complete.

Figure 8 shows data from one participant to illustrate the analysis method (group data shown in Results). The goal of the analysis was to determine, for each participant, the left-right spatial bias in judging line bisection. The X axis shows the 11 possible offsets of the horizontal line with respect to the transector, and the Y axis shows the proportion of trials that the

participant reported the line to be longer on the left. The data were fitted to a logistic function with a maximum and minimum fixed at 1 and 0 respectively. The midpoint of the logistic function, the point at which the curve crossed a height of 0.5, is indicated in the figure by a dotted vertical line. The X value of the midpoint was taken as a measure of the participant's left-right bias, the point at which the participant psychophysically judged the horizontal line to be bisected (the point of subjective equality). Following the standard interpretation of line bisection performance, if the horizontal line was longer on the left when the participant judged it to be bisected, then the participant was inferred to have a rightward spatial bias; likewise, if the horizontal line was longer on the right when the participant judged it to be bisected (as in Figure 8), then the participant was inferred to have a leftward spatial bias.

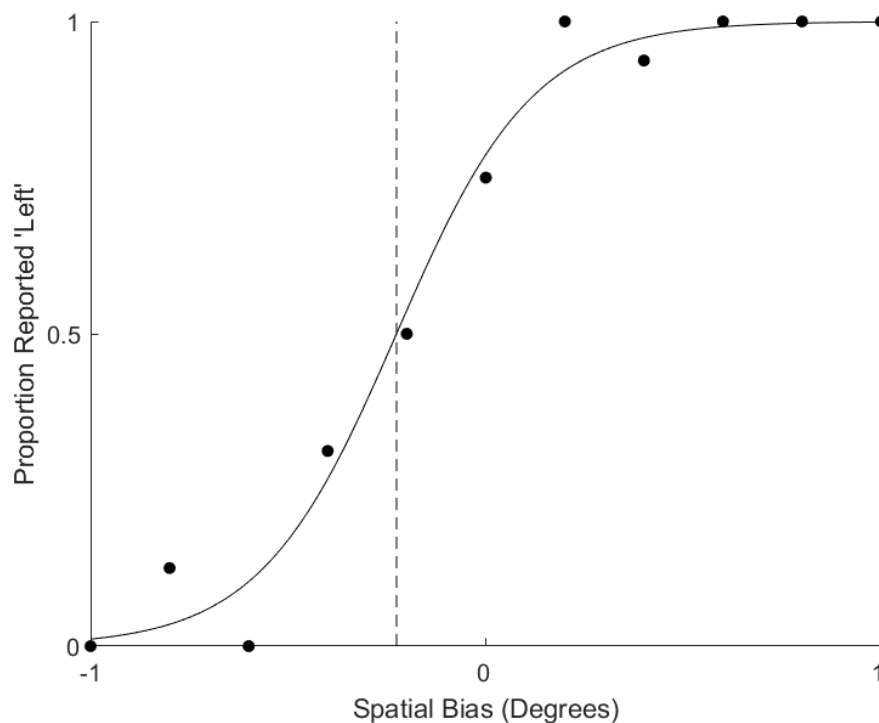


Figure 8. Results of the line bisection task for one participant to illustrate analysis method. x axis shows the visual degrees by which the horizontal line was longer on the left side of the transector. y axis shows the proportion of trials the participant reported the horizontal line to be longer on the left. The data were fitted with a logistic function. The dashed line shows the half-height point on the curve, the

psychophysical point at which the participant appeared to judge the line to be bisected. This participant had a negative, or leftward, spatial bias (the participant judged the line to be bisected when the right side was actually longer).

To provide an estimate of the quality of the curve fit for each participant, an r-squared value was computed. Only participants with an r-squared of 0.8 or higher were included in the final analysis, to increase the likelihood of obtaining a reliable measure of spatial bias for each individual participant. This criterion eliminated 21 of 79 participants, which we recognize to be a large exclusion, however, the central hypothesis of the study could not be tested without a reliable measure of spatial bias for each individual participant. A noisy or unreliable measure that failed to distinguish left from right pseudoneglect would render the hypothesized relationship between the two tasks difficult or impossible to detect.

Theory-of-Mind Task

Theory of mind is often tested with the false-belief task first introduced by Wimmer and Perner (1983). In the typical task, participants are presented with a vignette: Maxi helps his mother put groceries away by putting some chocolate in a blue cupboard. He then leaves for the playground. His mother uses the chocolate to make a cake and places the remaining chocolate in a green cupboard. When Maxi returns later to get some chocolate, where will he check for it? Solving the task requires an understanding that Maxi has a mind, that his mind contains beliefs about the world, and that he can hold beliefs that are contrary to fact (Baron-Cohen et al., 1985; Turner & Felisberti, 2017; Wellman et al., 2001; Wimmer & Perner, 1983).

In the modified task that we used, participants saw a cartoon that included people, boxes, and a ball. The ball was in one of two boxes and the participant had to decide whether a cartoon person would believe the ball to be in box 1, to the left of the cartoon person, or box 2, to the

right at the end of each trial. We used a reaction time measure to determine if the participant demonstrated a spatial processing bias toward the cartoon person's left or right. As detailed below, to convert the paradigm to a reaction time task, we presented participants with two cartoon people at the same time. Participants responded only at the end of the trial when one of the two cartoon people was indicated as the target for the theory-of-mind judgment.

Figure 9 shows the paradigm for the second task. Each trial began with a black fixation cross at the center of a white background. Participants were instructed to fixate on the cross. After 500 ms, the fixation cross was joined by a top-down view of two cartoon heads and two numbered boxes. The heads and boxes were arranged vertically with respect to the participant, such that the left-right axis with respect to the cartoon heads was orthogonal to the left-right axis with respect to the participant. In this way, the participant's own left-right bias could be measured (in the line bisection task) independently of a social cognition bias toward the left or right of the cartoon head.

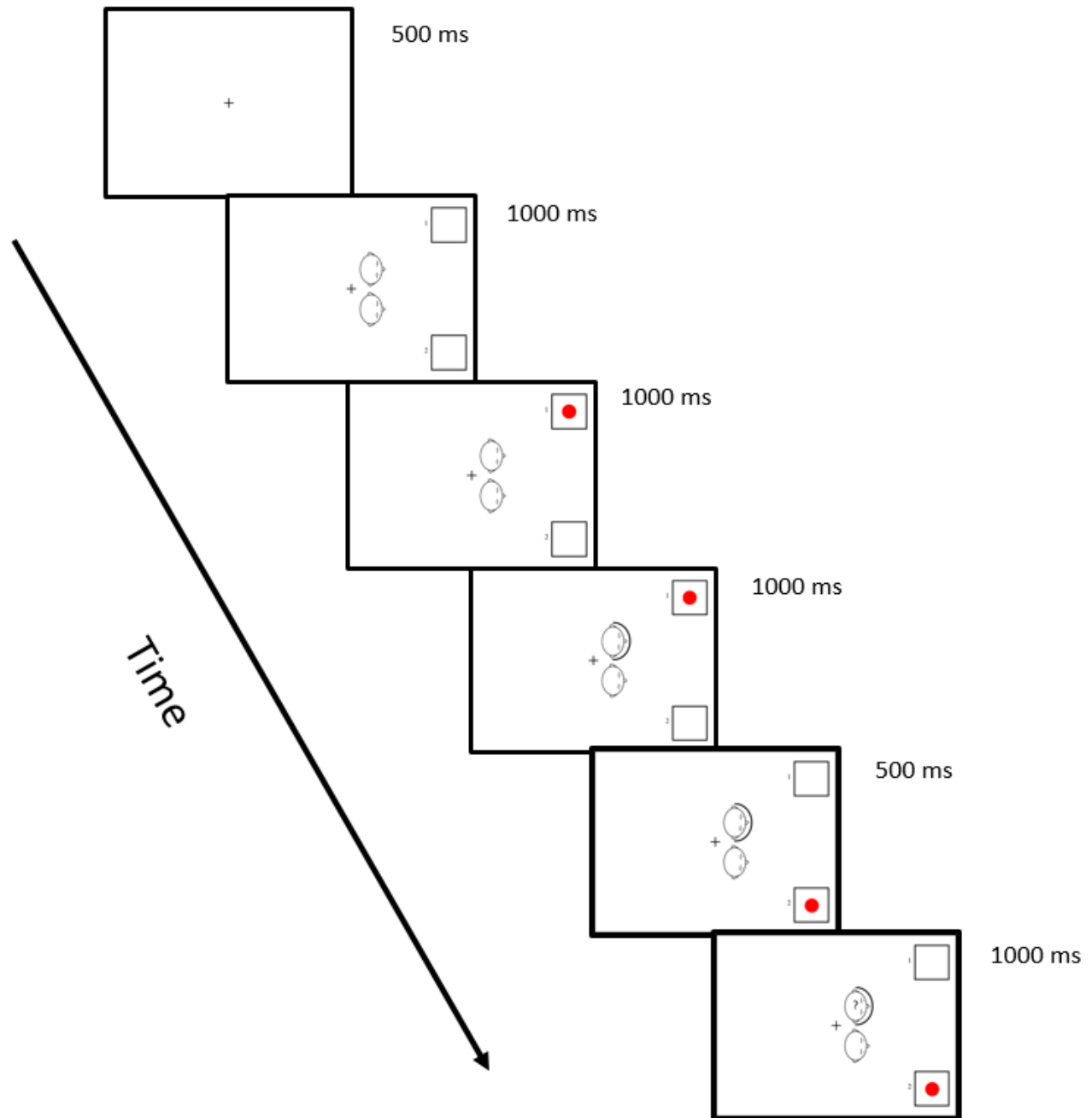


Figure 9. Paradigm for the theory-of-mind task. Participants saw a sequence of cartoon panels showing two heads viewing a ball that could be placed in either of two boxes. One of the heads had its vision blocked partway through the trial. In the final image, one head was indicated by a question mark. Participants were instructed to respond as quickly as possible, judging whether the indicated head would think the ball was in box 1 or box 2.

The heads were 5 degrees right of the vertical midline of the screen, facing toward the right, centered 5 degrees above and below the midpoint of the screen. The boxes were 23 degrees to the right of the midpoint of the screen, centered 15 degrees above and below the horizontal midline. After 1000 ms, a red ball appeared in one of the two boxes (half of the trials in box 1, half of the trials in box 2). Participants had been told in the instruction period that, in this configuration, both heads could see where the ball was located. After 1000 ms, one of the heads was blocked with a curved partition directly in front of it (half of the trials blocking the upper head, half of the trials blocking the lower head). Participants had been told in the instruction period that the blocked head could no longer see either the boxes or the ball, but that the other head could still see everything as before.

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

Finally, 4000 ms after the start of the trial, a question mark appeared inside one of the heads (half of trials in the upper head, half of trials in the lower head). The question mark indicated which head was to be the target of the participant’s judgment. The participant was instructed to respond as quickly as possible once the question mark appeared. By pressing one of two buttons on a standard keyboard, the participant reported whether the indicated head would believe the ball was in box 1 or box 2. Participants were given a response window of 1000 ms. Trials in which participants exceeded the given time to respond were not included in the analyses, though participants responded on most trials (95%). After the response window, the

display of heads and boxes disappeared, and the next trial began with the onset of the fixation cross.

Participants performed 256 trials, over 8 blocks of 32 trials each. The task took 20 to 30 minutes to complete. The task included the following conditions: the ball could be initially presented in box 1 or box 2; the blocking partition could be placed in front of the upper or lower head; the ball could be switched to the opposite box or remain in the same box; and the question mark could be presented in the upper or lower head. This 2X2X2X2 design resulted in 16 unique trial types, presented in a counterbalanced and randomized order.

Figure 10 illustrates the analysis by showing data from one participant (group data is discussed in Results). The trial types were collapsed into 4 conditions for purposes of analysis, 64 trials per condition. These conditions formed a 2X2 design as follows: the ball remained in the same box (“unswitched” trials) or switched to the opposite box (“switched” trials), and the head indicated by the question mark was the one not blocked by the screen (“unblocked” trials) or was the one blocked by the screen (“blocked” trials). For each of these four conditions, the correct answer could be box 1 (the box that was to the left of the head) or box 2 (the box that was to the right of the head). As shown in Figure 10 for one subject, a mean reaction time was computed for each of those conditions. A reaction time difference score was then computed (difference score = [mean reaction time for the left-box trials] – [mean reaction time for the right-box trials]). A positive score indicated a faster reaction time when the right box was the correct answer, and thus a right-sided preference. This difference score was computed for each of the four main conditions, for each participant. For example, in Figure 10, the participant had a right preference in the blocked-switched condition, but little or no preference in the other three conditions.

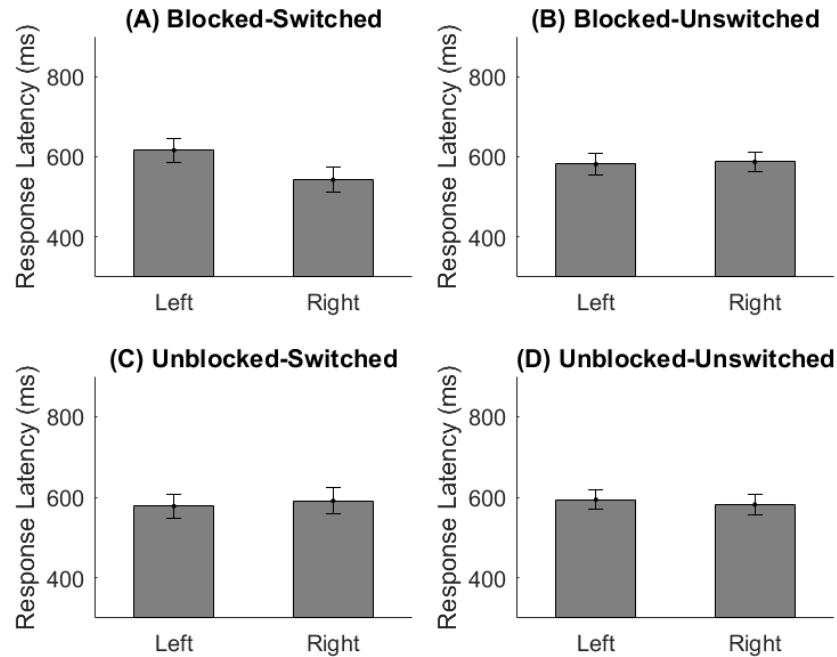


Figure 10. Result of the theory-of-mind task for one participant, to illustrate analysis method. y axis shows mean response time. The ball remained in the same box (unswitched trials) or switched to the opposite box (switched trials). The cartoon head indicated by the question mark was the one not blocked by the screen (unblocked trials) or was the one blocked by the screen (blocked trials). The correct answer could be the box that was to the left of the head (“left” trials) or the box that was to the right of the head (“right” trials). (A) This participant showed a latency difference in the blocked-switched condition, with an advantage in processing the right side with respect to the cartoon head. (B) No significant latency difference was found in the blocked-unswitched condition for this participant. (C) No significant latency difference was found in the unblocked-switched condition. (D) No significant latency difference was found in the unblocked-unswitched condition.

The blocked-switched condition corresponded to a false-belief condition (the participant could see the ball was in one box, but the cartoon character should “think” the ball is in the other box). The other three conditions corresponded to true-belief conditions (the participant could see the ball was in one box and the cartoon character should “think” the ball is in the same box).

Most participants performed at high accuracy (mean of 86% of trials correct). Only correct trials were included in the analysis. Three subjects were removed from analysis due to below-chance performance suggesting that they had not understood the instructions. Given this removal, and the removal of participants based on curve-fit criteria in the line bisection task, we

had a final set of 55 participants in which the performance on the two tasks could be compared (39 female, age 18-55 years, median age of 21).

Results

Line Bisection Task

A spatial bias was computed for each participant. As described in the Methods and illustrated in Figure 8, we used a logistic curve fit to obtain the point of subjective equality, the point at which the participant appeared to judge a horizontal line to be bisected by a vertical line. Following the standard interpretation of line bisection performance, if the point of perceptual bisection corresponded to a longer left side, the participant was inferred to have a spatial bias in favor of the right. If they more often judged the right side was longer, the participant was inferred to have a spatial bias in favor of the left. Figure 11 shows the distribution of spatial biases among 55 participants. Although each individual participant may have shown a spatial bias (such as the left-biased participant illustrated in Figure 8), the overall distribution was not significantly different from zero (mean = -0.03 degrees, SEM = 0.03, two tailed t test, $t = 1.22$, $df = 54$, $p = 0.220$), indicating that neither a right nor a left bias predominated this sample. This finding is not inconsistent with previous findings that the average spatial bias can be subtle and variable across studies (Jewell & McCourt, 2000). The focus of the present study was not on the average group bias, however, but on how individuals' spatial biases corresponded between the two tasks.

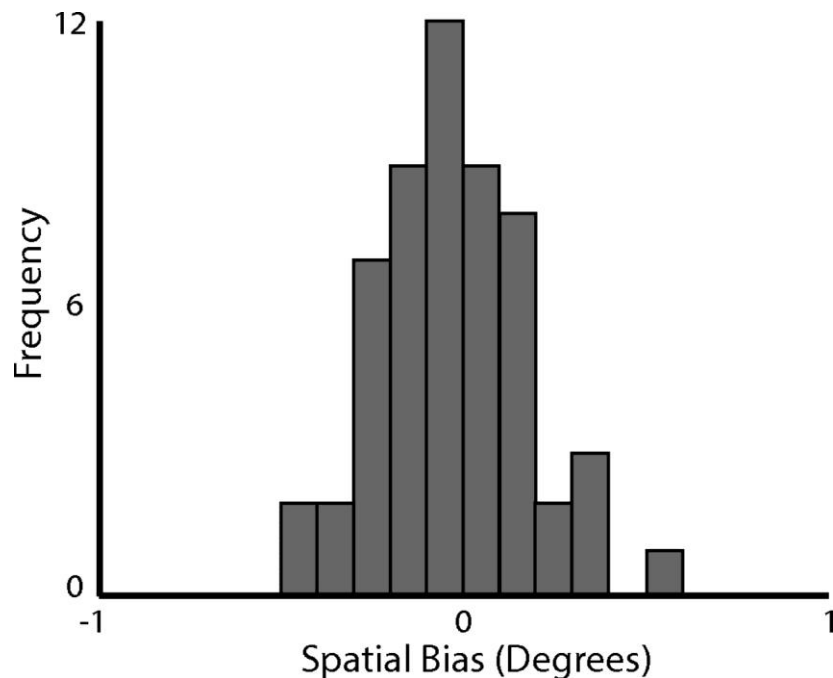


Figure 11. Frequency histogram showing the spatial bias, measured in the line bisection task, among 55 participants. Although individual participants showed a spatial bias, the sample as a whole showed no significant overall left or right bias (two-tailed t test, $t = 1.22$, $df = 54$, $P = 0.220$).

Theory-of-Mind Task

On each trial of the theory-of-mind task, a reaction time was measured. The mean reaction time for all trials was 593 ms (SEM = 10.5). As described in the Methods and illustrated in Figures 9 and 10, the task included four conditions corresponding to a 2X2 design. The blocked-switched condition corresponded to a false-belief condition (in which the cartoon character should “believe” the ball to be in one box when it is actually in another). The other three conditions were true-belief conditions (in which the cartoon character should “believe” the ball to be in the box that it is actually in). Table 1 shows the mean reaction times in these four conditions. An ANOVA indicated that there was a significantly greater reaction time for switched trials than for unswitched trials (within subjects 2X2 ANOVA, $F = 13.80$, $p = 0.0003$), presumably because of the extra cognitive cost of processing the switched ball location. There

was no significant difference between blocked and unblocked trials, and no significant interaction (main effect of blocked vs. unblocked, $F = 3.53$, $p = 0.062$; interaction, $F = 0.13$, $p = 0.714$). The focus of this experiment, however, is not on the group statistics but on how individual spatial biases corresponded between the two tasks.

Conditions	BS	BnS	nBS	nBnS
Mean	618.74	563.89	639.07	594.08
SEM	11.75	16.25	10.12	14.75

Table 1. Times shown in milliseconds. BS, blocked, switched condition; BnS, blocked, unswitched condition; nBS, unblocked, switched condition; nBnS, unblocked, unswitched condition.

In each of the four conditions in the theory-of-mind task, a reaction time difference score was obtained for each participant. This difference score is explained in detail in the Methods section. The difference score corresponded to a spatial bias, with a positive score corresponding to a rightward bias with respect to the cartoon head. The relationship between this spatial bias and the bias measured in the line bisection task is described in the next section.

Correspondence Between Tasks

Figure 12A shows how the spatial bias measured in the line bisection task relates to the spatial bias measured in the false-belief condition of the theory-of-mind task. The two measures are significantly, positively correlated (linear regression, $\text{Beta} = 0.77$, $F = 6.26$, $p = 0.016$), thus supporting the prediction.

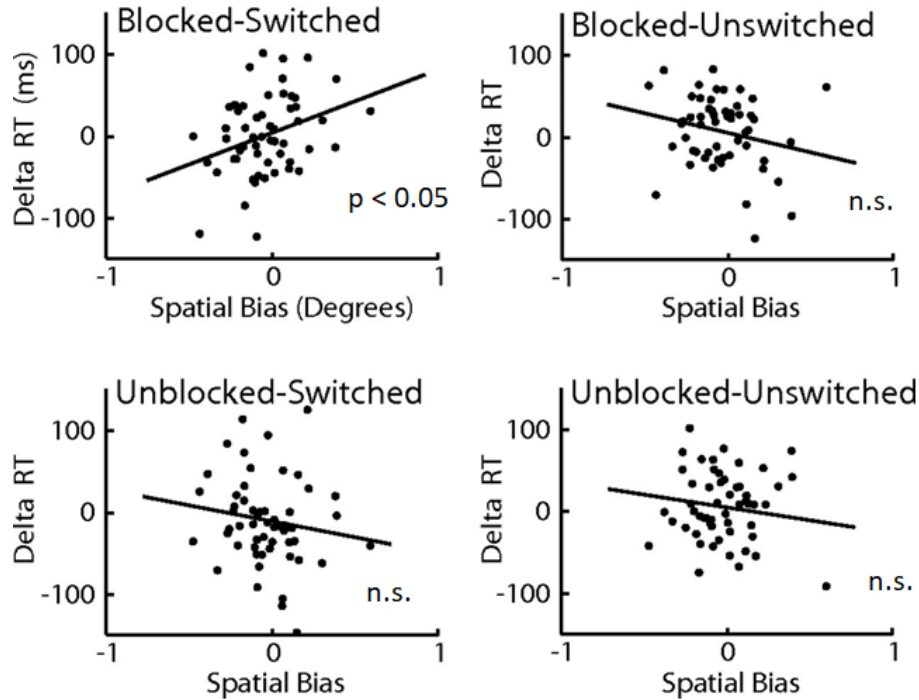


Figure 12. Regression plots relating the line bisection results to the theory-of-mind results in 55 participants. The x axis shows the spatial bias for each participant, measured in degrees of visual angle using the line bisection task. The y axis shows the participant's bias toward the left or right of the cartoon head, measured in the theory-of-mind task, using a difference of reaction times (delta RT). (A) Spatial bias versus theory-of-mind performance in the blocked-switched condition. Each point shows the results for one participant. The regression line is significant ($\text{Beta} = 0.77$, $F = 6.26$, $p = 0.016$). Because this condition was the only one to require true theory-of-mind reasoning, this regression is the crucial test of the hypothesis. (B–D) Similar plots for conditions in which theory-of-mind reasoning was not required to solve the task. These regression lines were not significant (B, $\text{Beta} = -0.48$, $F = 2.79$, $p = 0.101$; C, $\text{Beta} = -0.39$, $F = 1.20$, $p = 0.279$; D, $\text{Beta} = -0.31$, $F = 0.85$, $p = 0.360$).

As shown in Figure 12B, 12C, and 12D, in the three true-belief conditions that did not necessitate theory-of-mind reasoning, the correlations were nonsignificant and trended in a negative direction (unblocked-unswitched, $\text{Beta} = -0.48$, $F = 2.79$, $p = 0.101$; unblocked-switched, $\text{Beta} = -0.39$, $F = 1.20$, $p = 0.279$; blocked-unswitched, $\text{Beta} = -0.31$, $F = 0.85$, $p = 0.360$).

Discussion

The present study tested participants in two tasks: a line bisection task and a theory-of-mind task. The line bisection task measured whether each participant had a visuospatial bias toward the right or left. The theory-of-mind task measured whether each participant had a processing-speed bias along an orthogonal axis, toward the right of, or left of, the agent about which the participant was thinking. Performance on these two tasks was significantly correlated. The correlation was specific to the one condition in the theory-of-mind task that required social cognition, the false-belief condition. One possible interpretation is that the participants exhibited a spatial bias in the theory-of-mind task because, when social cognition was engaged, the participants were reconstructing a left-right frame of reference with respect to the cartoon character.

An alternative explanation for the observed relationship is that people might have both a horizontal and a vertical spatial bias, and a leftward bias might generally tend to correlate with an upward bias. This explanation is unlikely, given the previous finding that left-right and up-down pseudoneglect do not correlate (Heber et al., 2010). This alternative explanation is also inconsistent with the present findings, in which the relationship could be seen only for the false-belief condition in the theory-of-mind task, and not for the three true-belief conditions. The true-belief conditions did not show any significant correlation with the line bisection task. The results are more consistent with the interpretation that in the true-belief conditions, because the same level of social cognition was not necessary to solve the task, participants may have taken a cognitive short cut, or may have engaged social cognition to a lesser extent. In that case, on true-belief trials the participants might not have actively conceptualized the ball in relation to the cartoon head, and therefore their own left-right processing bias would not have been engaged.

The results are consistent with studies on perspective taking, in which people mentally rotate themselves into the spatial perspectives of others in order to judge the perceptual world of others (Blanke et al., 2005; David et al., 2006; Ruby & Decety, 2001). The results may also relate to the finding that reaction times in a theory-of-mind task are faster when the participant and the character share the same mental state (Kovács et al., 2010). One's own mental state, beliefs, or spatial processing limitations may be mirrored in how one attributes mind states to others.

The line bisection task was originally used to measure a severe spatial bias in the clinical syndrome of hemispatial neglect (Corbetta, 2014; Halligan et al., 2003; Heilman & Valenstein, 1979). In these cases, the patient's attention and awareness of one side of space is compromised. In neurotypical participants, the line bisection task is thought to measure a subtler spatial bias in attention to or awareness of the two sides of space (Brooks et al., 2014; Jewell & McCourt, 2000; Milner et al., 1992; Szczepanski & Kastner, 2013). In the theory-of-mind task in the present experiment, the participant is effectively attributing to the cartoon agent an attention to, or an awareness of, the ball in a specific location. We speculate that the mechanisms by which people direct their own spatial attention and awareness may partially overlap the mechanisms by which people attribute such mind states to others (Graziano, 2013; Graziano & Kastner, 2011). The cortical networks that appear to be involved in these different functions show substantial anatomical overlap, especially where they pass through the temporoparietal junction (TPJ). Damage to this region, especially on the right side, can lead to the most severe cases of hemispatial neglect (Vallar & Perani, 1986). Yet regions within the TPJ are also consistently recruited in studies of social cognition. Theory-of-mind tasks and perspective taking tasks activate at least some subregions within the TPJ (Blanke et al., 2005; Kelly et al., 2014; Saxe &

Wexler, 2005). A complex experimental literature suggests that the TPJ is composed of many functional subregions, with some separation and some overlap of function (Igelström et al., 2016; Igelström & Graziano, 2017; Mitchell, 2008; Scholz et al., 2009; Webb et al., 2016). Although the full extent of anatomical overlap and separation is not yet resolved, it appears that social, attentional, and awareness-related networks pass close to each other with partial overlap in the TPJ. Given this tight-knit anatomical and functional association, it is possible that this overlap contributes to the present findings in which one's own limitations in spatial processing are mirrored in one's social cognition.

Chapter 4 – Theory-of-mind deficits in patients with spatial neglect

Introduction

In the previous chapter we found evidence that an individual's own attentional bias transfers to the judgments they make about others' attention. This carryover of the way a person models the world around them to the way they model the world for others demonstrated that these processes are deeply entwined, if not based in the same neural mechanisms. In the study described in this chapter, we expand on the findings of Chapter 3, and address the question more directly. Building on the finding of transfer of attentional bias in neurotypical subjects, we asked whether stroke patients who were clinically assessed as suffering from hemispatial neglect showed a social cognition deficit, as measured by our version of a theory-of-mind task.

Hemispatial neglect is a clinical syndrome typically caused by damage to the right hemisphere, often to the parietal lobe or temporoparietal junction (Buxbaum et al., 2004; Heilman et al., 1987; Heilman & Valenstein, 1979; Parton et al., 2004). It manifests as a loss of attention to and subjective awareness of visual stimuli on the left side of the patient, and sometimes also extends to other sensory modalities. Essentially, it is the more severe analog to the pseudoneglect in neurotypical subjects we previously tested. To date, social cognition deficits have not been reported in association with hemispatial neglect.

Three possible outcomes were considered. In possibility 1, subjects have the ability for theory-of-mind reasoning. In that possibility, subjects should perform above chance level on all four conditions, including on the false-belief condition, which is typically thought of as relying most heavily on theory-of-mind reasoning. In possibility 2, subjects have a good cognitive ability to process and understand the instructions but have a deficit specific to theory-of-mind reasoning. In that possibility, subjects should perform above chance level on three true-belief

conditions but below chance level on the false-belief condition. A subject who consistently uses a non-theory-of-mind strategy would in principle earn a perfect score on the three true-belief conditions and answer every question wrong on the false belief trials. In possibility three, subjects have a general cognitive deficit, not specific to theory-of-mind reasoning. In that possibility, subjects should perform near chance level on all four conditions. Given the findings reported in Chapter 3, and the theoretical predictions of the Attention Schema Theory, we predict that patients suffering from hemispatial neglect, a disruption in one's awareness, should carry that deficit over to their modeling of others' awareness and therefore show disruption in social cognition displaying responses described in possibility 2.

One fundamental claim of the Attention Schema Theory is that our ability to model the attention of others relies on the same neural mechanisms that are used to construct a model of our own attention. In Chapter 3, we found evidence in a neurotypical population that supports that claim, by showing that a bias that effects one's own attention is also present when assessing the attention of another person. If patients suffering from neglect do indeed show difficulties with social cognition, it would be further support that there is overlap between neural areas that model attention and awareness for the self and others. For an individual suffering from a stroke that affects the temporoparietal junction causing loss of awareness, it is not immediately obvious that that individual should also experience additional deficits in their social cognition. AST however provides a framework of why this might be predicted, because constructing an agent's awareness, whether that agent be oneself or another person, are reliant on the same basic mechanisms.

Hemispatial neglect patients are a rare population and difficult to find. One challenge of working with this population is that hemispatial neglect can be transient. Patients assessed at the

acute timepoint, when they first incurred their injury or stroke, can be diagnosed with hemispatial neglect, but weeks or months later, often after rehabilitation when subjects are recruited, their neglect can be diminished. All subjects in the studies described in this chapter were recruited based on their initial neglect diagnosis at the time they were originally admitted. We acknowledge that it is difficult to know for certain the extent to which these subjects were experiencing neglect at the time that we tested them on social cognition.

Methods

Subjects

Subjects were recruited from the Casa Colina Hospital and Centers for Healthcare in Pomona, CA. All subjects provided informed consent and all procedures were approved by the Princeton and Casa Colina Institutional Review Boards.

Two groups of subjects were tested. Experiment 1 included five patients who had been clinically assessed (before recruitment into the study) as having left spatial neglect. Ages were 46, 29, 47, 48, and 72. Two of the five subjects were female.

Experiment 2 included a new set of five patients, again clinically assessed (before recruitment into the study) as having left spatial neglect. In addition, as part of the present study, the patients were tested with the Bells Test and Line Bisection Test at the time of the testing to confirm the presence of left spatial neglect. Ages were 37, 60, 22, 47, and 72. All 5 were male.

A follow up sample of age-matched control subjects is planned; however, those data are not available at the time of this composition. This set of control subjects will be recruited from a pool of individuals who have also suffered a stroke, but who have not at any time been diagnosed with hemispatial neglect. For both of the experimental groups as well as the control group, the exact location of the neural damage was not known due to a reliance on behavioral measures

rather than neural imaging. We still believe that despite lacking neural data and only having access to a modest sample size, the behavioral paradigm we employed is sufficient to reveal effects in this population.

Behavioral task

Figure 13 shows two versions of the task. The first version was more complex, and matched the paradigm used previously in neurotypical subjects described in Chapter 3. All visual components were positioned in the right half of the visual field, to allow patients suffering from left hemispatial neglect to see the stimuli. The heads were 5° right of the vertical midline of the screen, facing toward the right, centered 5° above and below the midpoint of the screen. The boxes were 20° to the right of the midpoint of the screen, centered 10° above and below the horizontal midline. In task 1 (Figure 13A), the following sequence of events occurred. First, two cartoon heads appeared for 1,000 ms, looking at two open boxes. Second, the “ball”, a red dot, appeared in one box for 1,000 ms. Participants had been told in the instruction period that, in this configuration, both heads could see where the ball was located. Third, one head was shown having its eyes covered with a curved partition for 1,000 ms, such that it could no longer see what was in either box, while the other head remained with uncovered eyes. In the instruction period, participants had been told that the blocked head could no longer see either the boxes or the ball but that the other head could still see everything as before. Fourth, on half of the trials, the ball switched from one box to the other. For both switched and unswitched trials, this stage of the trial lasted 1,000 ms. Fifth and finally, a question mark appeared over one head or the other. The question mark indicated which head was to be the target of the participant’s judgment. The participant was allowed a response window of 5,000 ms and instructed to respond as quickly as possible once the question mark appeared.

The task was to decide whether the head indicated by the question mark “believed” the ball to be in box 1 or box 2. Subjects responded by pressing the K or M key on a standard keyboard. The purpose of the incremental presentation of visual information was to ensure that the subjects would not know how to answer the question until the last piece of information, the question mark, was presented. At that moment, subjects had sufficient information to make a theory-of-mind judgment and responded in a speeded manner within a limited response window. In our previous study on neurotypical subjects, the response window was 1.5 s. Based on pilot results, we extended the response window in the clinical subjects to 5 s.

Group 2 (Figure 13B) used a simplified version of the task after the first version proved to be difficult for the clinical patients to navigate. Changes made to the task included reducing the number of cartoon agents on screen from two to one, modifying the placement and layout of stimuli on screen for clarity and changing the response buttons to the up and down arrow keys. The same general design, however, was preserved. In this version of the task, the fixation cross was removed, and the sole cartoon head was located at the horizontal midline of the screen. The boxes were “opened up” to show their contents more easily by removing the edge closest to the cartoon and they were both moved closer to the cartoon head. The boxes were 8° to the right of center, and 5.5° above and below the horizontal midline. As with the previous versions, the stimuli were confined exclusively to the right hemifield, and the same sequence of stages progressed.

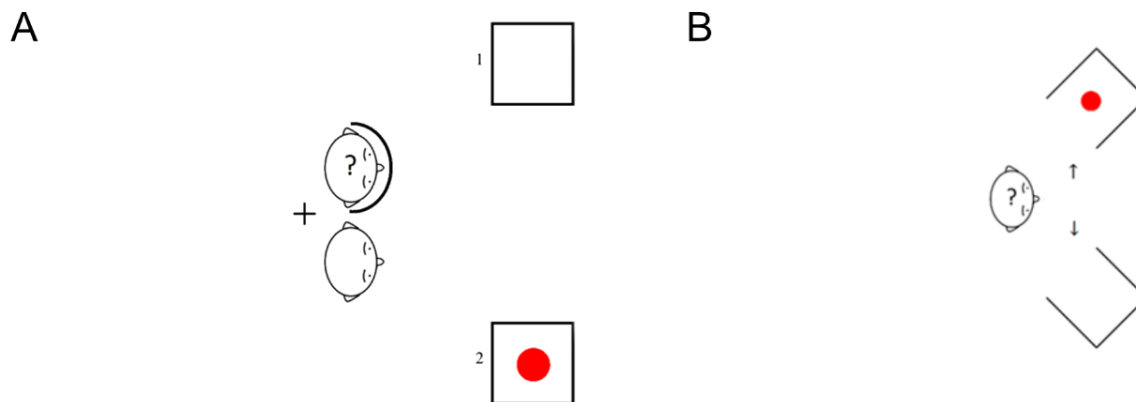


Figure 13. (A) Original theory-of-mind paradigm from Bio et al., 2018 used on neurotypical subjects and patients in experiment 1 in the present study. (B) Adapted paradigm used for patients in experiment 2.

For both versions of the task, we conceptualized the paradigm as a 2X2 design. The first variable was whether the cartoon head had its vision blocked: in some trials, the subjects had to judge the visual beliefs of a cartoon head whose eyes were covered, whereas in other trials, the subjects had to judge the visual beliefs of a cartoon head whose eyes were uncovered. The second variable was whether the ball switched from one box to the other: in some trials, the ball switched boxes midway through the trial, whereas in other trials, the ball began in one box and remained there without switching. This design resulted in four trial conditions that we termed blocked-switched (BS), blocked-unswitched (BnS), unblocked-switched (nBS), and unblocked-unswitched (nBnS). The significance of these four conditions is explained further in the Results.

Results

In experiment 1, five patients were tested with task 1, and the results were compared to the previously collected results from neurotypical subjects on the same task (from Bio et al., 2018). In the BS condition, the cartoon head “thinks” the ball is in one box when it is actually in the other box; thus, the cartoon has a false belief. To answer correctly, the subject must have the social cognitive ability to recognize that the cartoon has a mind, an awareness of conditions in its

environment, and that the mind can be deceived and have beliefs that are contrary to reality. In the other three conditions, the BnS, nBS, and nBnS conditions, the cartoon head “thinks” the ball is in the box where it actually is located; thus, the cartoon has a true belief. To solve these three true-belief conditions, the subject does not require any theory-of-mind ability, but can simply report which box the ball is actually in.

Figure 14A shows the results from the study presented in chapter 3 in which 79 neurotypical subjects were tested on task 1. These subjects clearly matched possibility 1 listed above. They performed above chance level (50% accuracy) on all four conditions (BS: $M = 76.45$, $SEM = 2.05$; BnS: $M = 92.56$, $SEM = 1.53$; nBS: $M = 81.71$, $SEM = 1.86$; nBnS: $M = 92.50$, $SEM = 1.47$). The lack of perfect scores on such a simple task is almost certainly the result of requiring a speeded response. Yet despite the time pressure, subjects performed well above chance on all conditions, including the false belief (BS) condition, indicating that they successfully used a theory-of-mind strategy.

In comparison, Figure 14B shows the performance on task 1 of five patients who were clinically assessed as suffering left spatial neglect. The subjects clearly matched possibility three. They showed a general cognitive deficit, scoring near chance level on all conditions. (BS: $M = 49.82$, $SEM = 14.53$; BnS: $M = 57.42$, $SEM = 10.47$; nBS: $M = 66.24$, $SEM = 9.77$; nBnS: $M = 73.34$, $SEM = 8.37$). They did not show a specific deficit on the false belief condition.

The results of experiment 1 do not support the hypothesis that hemispatial neglect patients have a specific deficit in theory-of-mind reasoning. However, they also do not rule out the hypothesis. The neglect patients evidently suffered enough general cognitive confusion that they could not perform above chance level in any condition in task 1. We therefore modified the task, reducing its level of difficulty. Task 2 was a simplified version of task 1, designed to be

easier for clinical patients to perform. If social cognition and spatial neglect are not linked, then with the easier task, patients should show a general, improved, above-chance performance on all four conditions. In contrast, if poor social cognition is specifically associated with spatial neglect, then in the easier task, patients should show an improved, above-chance performance only on the three true-belief conditions, and should show a reduced, below-chance performance on the false belief condition.

Figure 14C shows the results. The subjects clearly matched possibility 2. They performed above chance level on the three true-belief conditions, and below chance on the false belief condition (BS: $M = 18.37$, $SEM = 17.59$; BnS: $M = 97.48$, $SEM = 1.19$; nBS: $M = 98.13$, $SEM = 1.88$; nBnS: $M = 99.37$, $SEM = 0.39$). This pattern indicates enough cognitive sophistication to understand and solve the task, but also indicates the presence of a specific deficit in theory-of-mind reasoning. By applying a non-theory-of-mind strategy across all conditions, subjects performed well in the true-belief conditions and answered almost every question wrong in the false-belief condition.

The results from both experiments suggest that, while patients with hemispatial neglect may suffer some general cognitive deficits (as indicated by their across-the-board poor performance in task 1), when the task is made easier, a more specific deficit emerges. The patients show a specific social cognition deficit, as indicated by their inability to solve the theory-of-mind task.

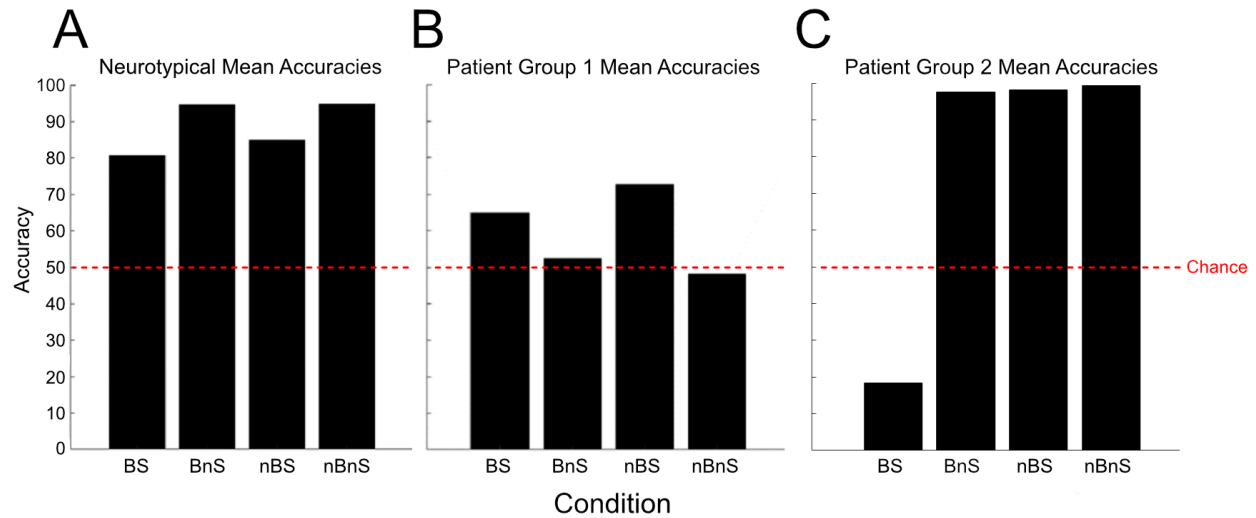


Figure 14. Average group results on the theory-of-mind task. There are 4 distinct conditions: a false-belief condition (blocked-switch, BS), and three true-belief conditions (blocked-unswitched, BnS; unblocked-switch, nBS; unblocked-unswitched, nBnS). A) Neurotypical subjects B) Pilot patient group with original task format C) Patient group with altered task format.

Discussion

The present study asked whether patients who were clinically assessed as having a left spatial neglect might show a specific deficit in theory-of-mind reasoning. To assess theory-of-mind reasoning, we used a visual form of the false belief task, in which four different conditions were presented. One condition, the false belief condition, required the subjects to use a theory-of-mind strategy. In the remaining three conditions, the true belief conditions, a successful answer did not require theory-of-mind reasoning. We found in a previous study that neurotypical subjects performed above chance level on all four conditions (Bio et al., 2018).

In the present study, in experiment 1, five subjects diagnosed with a left spatial neglect showed a general deficit across all four conditions, performing near chance levels, suggesting a general cognitive impairment. In experiment 2, the task was simplified and slowed, to try to compensate for the general cognitive impairment found in experiment 1. We tested five new patients who were clinically assessed as having left spatial neglect. With the simplified task,

performance improved on the three true belief conditions, and was near ceiling. At the same time, performance plummeted below chance on the false belief condition. It is telling that subjects did not merely perform near chance level on the false belief condition, but performed well below chance, approaching zero percent correct, a score that can be explained only if the subjects consistently used a non-theory-of-mind strategy. The results of experiment 2 therefore suggest that patients with spatial neglect are especially impaired at social cognitive thinking. They may have a general cognitive deficit, as indicated in experiment 1, but when they were given the easier task 2, a specific social cognitive deficit emerged.

Although these findings are the first of their kind, they are predicted by the Attention Schema Theory. The theory proposes that a key component of understanding the minds of others, social cognition, is modeling others' attention. Because AST also states that this process is reliant on the same mechanisms as those that are responsible for modeling one's own attention, patients suffering from a disruption in their ability to properly model their own attention would naturally also struggle carrying that process out for others.

It is important to recognize that the sample size for this study was fairly modest, so although the effects were robust, the results should be interpreted with that as a consideration. That being said, the findings in this study shed light on an issue those suffering from neglect have not previously had acknowledged. While other mental health disorders such as autism and schizophrenia have had a large focus on the social deficits that accompany these ailments, visual neglect is sorely lacking any investigation into these issues. This work demonstrates the importance of continuing research in this area. More work should be done to explore the effect hemispatial neglect has on social cognition, as well as the kinds of interventions that could be beneficial in mitigating these problems.

Chapter 5 - Encoding the mental experience of others versus the truth-value of beliefs

Introduction

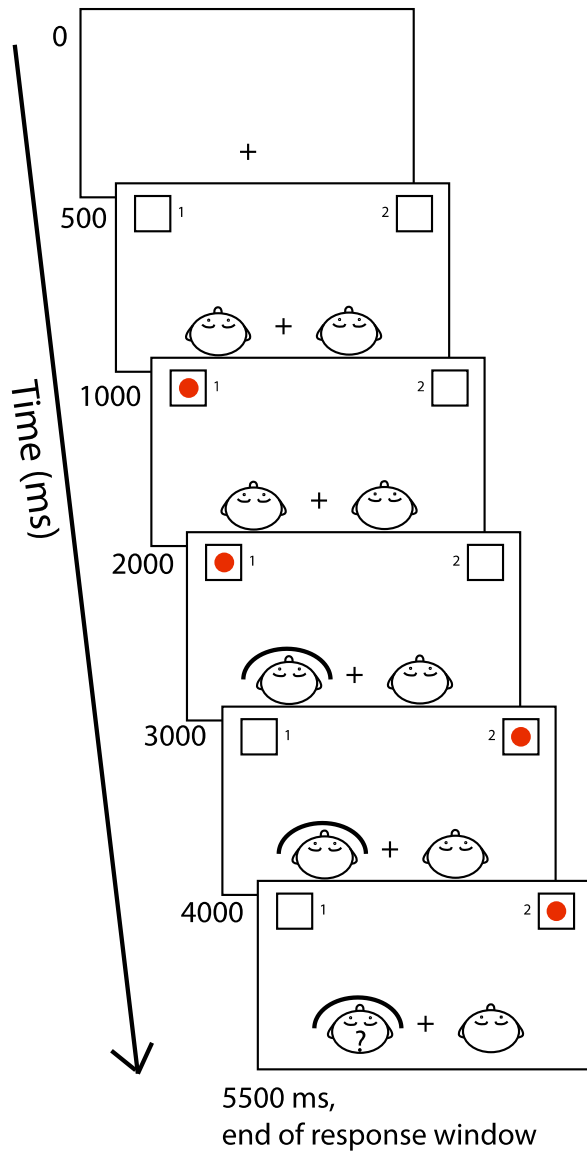
Building a model of other people's thoughts, emotions, and beliefs, also called theory of mind, is foundational to our social lives (Baron-Cohen, 1997; Frith & Frith, 2003; Wellman, 2018; Wimmer & Perner, 1983). A large literature shows that ToM tasks tend to activate a specific network of areas in the human cerebral cortex (e.g., Fletcher et al., 1995; Frith & Frith, 2006; Gallagher & Frith, 2003; Gallagher et al., 2000; Gobbini et al., 2007; Igelström et al., 2016; Kelly et al., 2014; Mar, 2011; Richardson et al., 2018; Saxe, 2006; Saxe & Kanwisher, 2003; Schurz et al., 2014; Spreng et al., 2009; van Veluw & Chance, 2014; Vogeley et al., 2001). Those areas typically include the temporoparietal junction (TPJ), sometimes bilaterally but with a bias toward the right side; the superior temporal sulcus (STS), again sometimes bilaterally but with a bias toward the right side; the medial prefrontal cortex (MPFC); and the precuneus. Other brain areas are also reported in ToM studies, such as the temporal pole and the amygdala, but the areas listed above are often more consistently active during ToM tasks, as shown in meta-analysis studies (Mar, 2011; Schurz et al., 2014; van Veluw & Chance, 2014). This chapter describes an MRI experiment on human subjects, which is currently submitted under consideration as *Right Temporoparietal Junction Encodes Mental Experience of Others More Than Evaluating False Versus True Belief* (Bio et al., 2021).

Experiments on the ToM cortical network often use a classic paradigm called the false belief task, which was briefly reviewed above in Chapter 1 (Baron-Cohen et al., 1985; Wimmer & Perner, 1983). In it, the subject of the experiment must decide, based on information known to be available to a character in a story, whether the character thinks that A or B is true. For

example, if Sally originally put her sandwich into box A, will she still think it is in box A, even after someone else, unbeknownst to Sally, has moved it to box B? To answer correctly, the subject of the experiment must have a sophisticated enough theory of Sally's mind to realize that Sally can believe something that is false – that is distinct from the world around her – depending on the information available to her.

In a recent behavioral study, described in Chapter 3, we designed a variant of the false belief task (Bio et al., 2018). Our version was designed to build up a set of information visually over the course of each trial. Visual ToM tasks incorporating cartoons, or, more often, videos, have been used before (e.g., Gallagher et al., 2000; Grezes et al, 2004; Hooker et al., 2008; Richardson et al, 2018; Rothmayr et al, 2011; Sommer et al., 2010; Marjoram et al., 2006). The experimental design is shown in Figure 15A, which is adapted from the paradigm described in Chapter 3 (Bio et al., 2018). This design was chosen, because it converts the typical false-belief task into a visual, reaction-time task, in which a single event (the appearance of the question mark) triggered the moment when subjects needed to make a ToM judgment. With all stimuli being fully right-versus-left counterbalanced across trials, the conditions were nearly visually identical, providing a good control for each other. Two small adjustments were made to the previous paradigm to make it more suitable for use in the MR environment; the configuration was rotated 90° counterclockwise to make all the items on-screen more easily visible, and an intertrial interval from 3-5 seconds was added to make recording signal clearer.

A Example Trial



B Main Trial Conditions

		Switched	non Switched
	Blocked	BS	BnS
		Contrast to test Hypothesis 1	
	non Blocked	nBS	nBnS
		Contrast to test Hypothesis 2	

Figure 15. Behavioral paradigm. (A) Timeline of events during a typical trial. Fixation point appeared at start of trial. Then two heads and two boxes appeared. Then a red ball appeared in one box. Then one head had its sight blocked by the curved barricade. Then, on half of trials, the ball switched to the opposite box. Then a question mark appeared in one head, signaling subjects to respond by deciding whether the indicated head “thinks” the ball is in box 1 or box 2. All events were right-left counterbalanced among trials. (B) Four main trial conditions formed by the 2 X 2 design of blocked versus nonblocked configurations X switched versus unswitched configurations. BS and BnS trials were contrasted to test hypothesis 1, whereas nBS and nBnS trials were contrasted to test hypothesis 2.

We conceptualized the paradigm as a 2X2 design (see Figure 15B). The first variable was whether the indicated cartoon head had its vision blocked: in some trials, the subjects had to judge the visual beliefs of the cartoon head whose eyes were covered, whereas in other trials, the subjects had to judge the visual beliefs of the cartoon head whose eyes were uncovered. The second variable was whether the ball switched from one box to the other: in some trials, the ball switched boxes midway through the trial, whereas in other trials, the ball began in one box and remained there without switching. This design resulted in four trial conditions that we termed blocked-switched (BS), blocked-unswitched (BnS), unblocked-switched (nBS), and unblocked-unswitched (nBnS).

In the present experiment, subjects performed the reaction-time, ToM task in a magnetic resonance imaging (MRI) scanner, which measured brain activity evoked by the subjects' ToM decisions on each trial. The experiment differed from many previous brain imaging ToM experiments in two major ways. First, many previous experiments used a block design, relying on an average of brain activity across many seconds as subjects performed a continuous task such as reading a story (e.g., Fletcher et al., 1995; Gallagher et al., 2000; Gobbini et al., 2007; Lee et al., 2011; Saxe & Kanwisher, 2003; Vogeley et al., 2001). In the present experiment, we analyzed brain activity evoked by a decision that occurred within a brief, reaction-time window. This focus on a narrow decision window was expected to result in a weaker brain activity signal, since it was averaged over a smaller time interval. We hoped, however, that focusing on a distinct moment of decision might increase the specificity of the results by reducing the influence of other aspects of the trial that were the same across conditions.

Second, many previous ToM experiments tested a contrast between a ToM condition, such as a false-belief condition, and a nonsocial control, such as a false photo, false sign, or

physical description condition (Fletcher et al., 1995; Gallagher et al., 2000; Gobbini et al., 2007; Lee et al., 2011; Richardson et al., 2018; Saxe & Kanwisher, 2003; Vogeley et al., 2001). Such designs are especially effective at revealing the cortical ToM network. However, many incidental, unintended differences might be present between social and nonsocial tasks, potentially complicating the interpretation of results. Here, we examined differences among ToM conditions that were visually and cognitively nearly identical, but that involved slightly different ToM decision processes at the crucial moment during the trial. We hoped that this design would help test more specific hypotheses about the cortical ToM network.

Although many different comparisons are possible among the four conditions (and are shown in the *Post-hoc analyses* section below), in this chapter the focus is on two specific hypotheses. In hypothesis 1, brain areas in the ToM network should respond differently to false belief and true belief trials. Many previous studies have compared false-belief to true-belief conditions, to test whether false-belief reasoning requires especially complex or intensive ToM or whether it is processed in a specific part of the ToM network as distinct from true-belief reasoning (Aichhorn et al., 2009; Döhnelt et al., 2012; Hooker et al., 2008; Sommer et al., 2010). The results of these previous studies are mixed. Though there is evidence of false-belief processing emphasized in some subregions of the ToM network, especially in the TPJ and the MPFC (Aichhorn et al., 2009; Döhnelt et al., 2012; Sommer et al., 2010), some researchers have argued that false and true-belief conditions are similar in how they recruit ToM reasoning (Saxe & Kanwisher, 2003). Here, to help address this ongoing question, we tested whether false-belief trials and true-belief trials affect ToM brain areas differently. The BS trials correspond to the only false-belief condition in our task. The head sees the ball in one box, then the head's eyes are covered, then the ball is switched to the other box. To answer correctly, subjects must realize that

the head “believes” the ball to be in the first box, even though the ball is actually in the second box. In contrast, the BnS trials represent a matching true-belief condition. In the BnS trials, once again the head sees the ball in one box and then the head’s eyes are covered. The ball, however, is not switched to the opposite box, and therefore, by the end of the trial, the head should believe the ball to be in the box in which it actually resides. The two trial types are closely matched except that one is a false-belief condition and the other is a true-belief condition. If hypothesis 1 is correct, then at least some cortical areas known to be involved in ToM should show significant activity in the comparison between BS and BnS trial types.

We proposed hypothesis 2 as an alternative to hypothesis 1. In hypothesis 2, the ToM brain areas build models of the cartoon character’s mind rather than tracking the truth value of the cartoon character’s belief. The crucial factor affecting the activity in the ToM brain areas, therefore, is not broadly what the subject knows, but only narrowly what the subject thinks the cartoon character knows. If the cartoon character is supposed to know the same information in two conditions – i.e., have the same mental experience – then the ToM brain areas should not distinguish between those conditions. If the cartoon character is supposed to know different information in two conditions – i.e., have different mental experiences – then the ToM brain areas should react differently to those conditions. In the BS and the BnS trials (the false-belief and true-belief conditions contrasted to test hypothesis 1), the cartoon head has the same implied mental state. The configuration, in both trial types, implies that the head “sees” the ball in a box, the head’s eyes are covered, and the head “believes” that the ball is in the same box in which it most recently saw it to be. In the perspective of the head, the two conditions are identical, and therefore the mental state that must be attributed to the head should be identical. Brain areas involved in reconstructing the mental states of others might therefore react in the same way to

both conditions. The subject may know that the character has a false belief in one condition (BS) and a true belief in the other condition (BnS), but the cartoon character should not know that. One suggestion (Mitchell, 2008) is that parts of the ToM network, especially the right TPJ, are involved in filtering out or ignoring what is actually happening right now in one's own experience and instead building a model of what could, hypothetically, be happening in another mind.

In hypothesis 2, the crucial difference between switched and unswitched trials, affecting activity in the cortical ToM network, should not emerge when the cartoon's eyes are covered (when the character cannot see the switch take place), but instead should emerge when the cartoon's eyes are uncovered (when the character can see the switch take place). Hypothesis 2 therefore predicts that the comparison between BS and BnS trials should not result in significant activity in the ToM network, because subjects are imputing the same mental experiences to the cartoon head; but the comparison between nBS and nBnS trials should result in significant activity in the ToM network, because subjects are imputing different mental experiences to the cartoon head. In nBS trials, by implication, the head might be thinking something like, "I first saw the ball there, then it switched, and now I know it's currently here instead." In nBnS trials, by implication, the head might think something simpler such as, "The ball is where I see it to be." To summarize, hypothesis 1 suggests that the crucial factor affecting ToM brain areas is what the subject knows about the cartoon character's beliefs (false belief versus true belief); whereas hypothesis 2 suggests that the crucial factor affecting ToM brain areas is specifically what the subject thinks the cartoon character knows.

There is a precedent for findings that would favor hypothesis 1 to explain what is going on in this kind of task (Kovacs et al., 2014; Saxe & Kanwisher, 2003; Saxe et al., 2006;

Schneider et al., 2014; Sommer et al., 2007). Two factors may explain these previous findings. First, each of these studies used either written stories or pictorial vignettes to compare situations where target agents possessed either false or true beliefs. While these methods were successful in evoking the desired epistemic state, there were many additional differences between conditions that did not have to do with the target agent's mental state. Second, each of these fMRI studies were performed with a block design and analyzed using univariate contrasts. Both of these approaches, while capable of detecting coarse differences such as epistemic states, cannot capture more subtle differences between mental states, such as thinking about the same environment with one physical change.

While hypothesis 1 is reasonable given the past literature, AST may predict that hypothesis 2 could actually be more likely. Although hypothesis 1 is not inconsistent with AST, hypothesis 1 is concerned more with the subject's knowledge, than the target agent's, and so, if one is interested in theory of mind, this metaknowledge – information about the truth value of the target agent's belief, rather than the belief itself, is less relevant information. AST posits that individuals create rich models of others' mental states, and this is exactly what is predicted in hypothesis 2. Here, a difference should be detected when comparing conditions with different mental states in the target individual. This paradigm was specifically chosen because of the tight controls between condition types to prevent the concerns of factor 1 listed above and uses more sensitive measures of analysis to avoid factor 2. Given the high level of control in the design, the more sensitive analysis methods, and the predictions of AST, we would expect hypothesis 2 to hold true in ToM regions.

To carry out the comparisons, we used multi-voxel pattern analysis (MVPA). The method trains a neural network classifier to decode, on the basis of the MRI signal, whether one or

another condition was presented on any given trial. If decoding is significantly above chance for a cluster of voxels in the brain, then that cluster contains information that differentiates the two conditions of interest. MVPA is considered to be more sensitive to the information content in voxels than more traditional subtraction methods. We used two MVPA approaches. First, for the main analysis, we used a targeted approach, testing specific ToM brain areas. We defined six regions of interest (ROIs) based on a previous meta-analysis of brain activity evoked during ToM tasks (van Veluw & Chance, 2014), in accordance with generally accepted guidelines in ROI analysis (Poldrack, 2007). These six areas included the TPJ bilaterally, STS bilaterally, precuneus, and MPFC (see Methods for definitions of ROIs). Second, we used a searchlight analysis that tested the whole brain for clusters of significant decoding, statistically corrected for multiple comparisons across the brain. The searchlight analysis is less statistically sensitive and was used here as an exploratory method to test whether any areas of the brain might have shown especially strong decoding outside the ROIs.

Methods

Subjects

All subjects provided informed consent and all procedures were approved by the Princeton Institutional Review Board. We tested 32 healthy human volunteers (17 females, 31 righthanded, aged 18-50, normal or corrected to normal vision). Four were removed from analysis due to poor MRI signal quality. Therefore, 28 subjects remained in the final analysis. Subjects were recruited from a paid subject pool, receiving 40 USD for participation.

Experimental setup

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

Behavioral task

The task is illustrated in Figure 15A. Participants saw a cartoon that included people, boxes, and a ball. The ball was located in one of two boxes and the participant had to decide whether a cartoon person would most likely believe the ball to be in box 1, to the left, or box 2, to the right. Participants responded only at the end of the trial when one of the two cartoon people was indicated as the target for the ToM judgment. Each trial began with a black fixation cross at the center of a white background. Participants were instructed to fixate on the cross. After 500 ms, the fixation cross was joined by a top-down view of two cartoon heads and two numbered boxes. The heads were centered 3.25 degrees to the left (head 1) and right (head 2) of the vertical midline of the screen, positioned on the horizontal midline (at the same height as the fixation cross). The boxes were centered 12 degrees to the left (box 1) and right (box 2) of the midline, and 9 degrees above the horizontal midline. After another 500 ms, a red ball appeared in

one of the two boxes (half of the trials in box 1, half of the trials in box 2). Participants had been told in the instruction period that, in this configuration, both heads could see where the ball was located. After 1000 ms, one of the heads was blocked with a curved partition directly in front of it (half of the trials blocking head 1, half of the trials blocking head 2). Participants had been told in the instruction period that the blocked head could no longer see either the boxes or the ball, but that the other head could still see everything as before.

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

Finally, 4000 ms after the start of the trial, a question mark appeared inside one of the heads (half of trials in head 1, half of trials in head 2). The question mark indicated which head was to be the target of the participant’s judgment. The participant was instructed to respond as quickly as possible once the question mark appeared. By pressing one of two buttons on the button box, the participant reported whether the indicated head would most likely think the ball was in box 1 or box 2. Participants were allowed a response window of 1500 ms. Trials on which participants exceeded the given time to respond were not included in the analysis. Participants responded within the correct time window on most trials (98%). After the response window, the display of heads and boxes disappeared and a variable, 1000 - 3000 ms inter-trial interval followed, after which the next trial began with the onset of the fixation cross.

In summary, the task included the following conditions: the red dot could be initially presented in box 1 or box 2; the blocking screen could be placed in front of the left or right head;

the red dot could be switched to the opposite box or remain in the same box; and the question mark could be presented in the left or right head. This 2X2X2X2 design resulted in 16 trial types, presented in a counterbalanced and randomized order. The trial types were collapsed into four main conditions for purposes of analysis. These conditions formed a 2X2 design as follows: blocked trials, on which the head indicated by the question mark was blocked by the screen, versus unblocked trials, on which the indicated head was not blocked by the screen; and switched trials, on which the ball moved to the opposite box, versus unswitched trials, on which the ball remained in the initial box. These four conditions were labeled as blocked switched (BS), blocked unswitched (BnS), unblocked switched (nBS), and unblocked unswitched (nBnS). Only the BS trials corresponded to a false belief condition, in which the indicated head should “believe” the ball to be in one box when it was actually in the opposite box. The other three conditions corresponded to true-belief trials, in which the indicated head should think the ball was in the box that it actually was in.

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

fMRI data acquisition

Functional imaging data were collected using a Siemens Skyra 3T scanner equipped with a 64-channel head coil. Gradient-echo T2*-weighted echo-planar images (EPI) with blood-oxygen dependent (BOLD) contrast were used as an index of brain activity (Logothetis et al., 2001). Functional image volumes were composed of 46 near-axial slices with a thickness of 3.0 mm (with no interslice gap), which ensured that the entire brain excluding the cerebellum was within the field-of-view in all subjects (80 x 80 matrix, 2.5 mm x 2.5 mm in-plane resolution, TE

= 30 ms, flip angle = 75°). Simultaneous multi-slice (SMS) imaging was used (SMS factor = 2). One complete volume was collected every 2 s (TR = 2000 ms). A total of 1300 functional volumes were collected for each participant, divided into 8 runs (130 volumes per run). The first three volumes of each run were discarded to account for non-steady-state magnetization. A high-resolution structural image was acquired for each participant at the end of the experiment (3D MPRAGE sequence, voxel size = 1 mm isotropic, FOV = 256 mm, 176 slices, TR = 2300 ms, TE = 2.96 ms, TI = 1000 ms, flip angle = 9°, iPAT GRAPPA = 2). At the end of each scanning session, matching spin echo EPI pairs (anterior-to-posterior and posterior-to-anterior) were acquired for blip-up/blip-down field map correction.

FMRI preprocessing

Results included in this manuscript come from preprocessing performed using FM RIPREP version 1.2.3 (Esteban et al., 2019), a Nipype (Gorgolewski et al., 2011) based tool. Each T1w (T1-weighted) volume was corrected for intensity nonuniformity using N4BiasFieldCorrection v2.1.0 (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh v2.1.0 (OASIS template). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009) was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0 (Avants et al., 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (Zhang et al., 2001) from the FSL suite (v5.0.9).

Functional data was slice time corrected using 3dTshift from AFNI v16.2.07 (Cox, 1996) and motion corrected using mcflirt (FSL v5.0.9) (Jenkinson et al., 2002). This procedure was followed by co-registration to the corresponding T1w using boundary-based registration (Greve

& Fischl, 2009) with six degrees of freedom, using flirt (FSL). Motion correcting transformations, BOLD-to-T1w transformation and T1w-to-template Montreal Neurological Institute (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation. Many internal operations of FMRIPREP use Nilearn (Abraham et al., 2014) principally within the BOLD-processing workflow.

Multivariate Pattern Analysis

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

The ROIs were defined as spheres centered on the statistical peaks reported in an activation likelihood estimation meta-analysis of 16 fMRI studies (including 291 subjects) involving ToM reasoning (van Veluw & Chance, 2014), in accordance with generally accepted guidelines in ROI analysis (Poldrack, 2007). The ROIs are shown in Figure 16. The peaks were located in six areas: the left TPJ (Montreal Neurological Institute [MNI]: -52, -56, 24), right TPJ (MNI: 55, -53, 24), left STS (MNI: -59, -26, -9), right STS (MNI: 59, -18, -17), MPFC (MNI: 1, 58, 19), and the precuneus (MNI: -3, -56, 37). The radius of the ROI spheres was 10 mm, corresponding to the approximate volume (4,000 mm³) of the largest clusters (TPJ and MPFC) reported in van Veluw and Chance (2014). The same sphere radius was used for all ROIs.

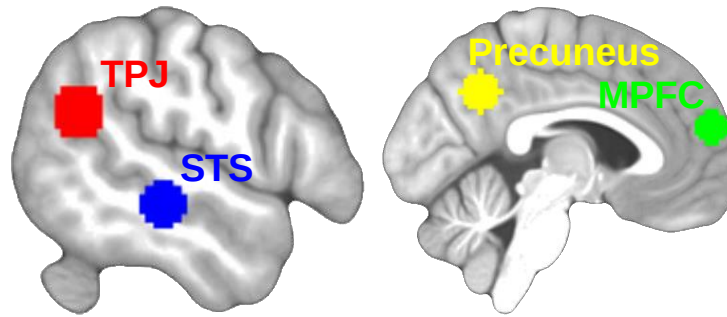


Figure 16. Regions of interest (ROIs). Six ROIs were defined based on peaks reported in an activation likelihood estimation meta-analysis of 16 fMRI studies involving theory-of mind reasoning. The ROIs consisted of 10-mm-radius spheres centered on peaks in the bilateral temporoparietal junction (TPJ)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

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

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

independent training and testing data sets, we used a leave-one-run-out cross-validation approach. For each fold, the betas across all training runs were normalized relative to the mean and standard deviation, and the same Z-transformation was applied to the betas in the left-out test run (Misaki et al., 2010). An SVM was then trained to discriminate activity patterns belonging to the contrasted trial types in seven runs, and then tested on the left-out run, repeated for all runs, resulting in a run-average decoding accuracy for each ROI and subject.

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

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

searchlight analysis is useful for revealing clusters of strong decoding in areas that were not anticipated by a hypothesis.

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

Results

Task performance

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

ROI analysis, Hypothesis 1

Figure 17 shows the results of the ROI analysis for hypothesis 1, in which false belief trials (BS) were compared with the corresponding true belief trials (BnS). None of the six ROIs showed any significant decoding (the decoding accuracy was not significantly different from the chance level of 50%). We did not find any evidence of a difference between processing false belief and true belief trials, within the six defined ROIs in known ToM cortical areas. Thus, hypothesis 1 was not supported by the analysis within the six ROIs. Statistics for tests run can be found in table 2.

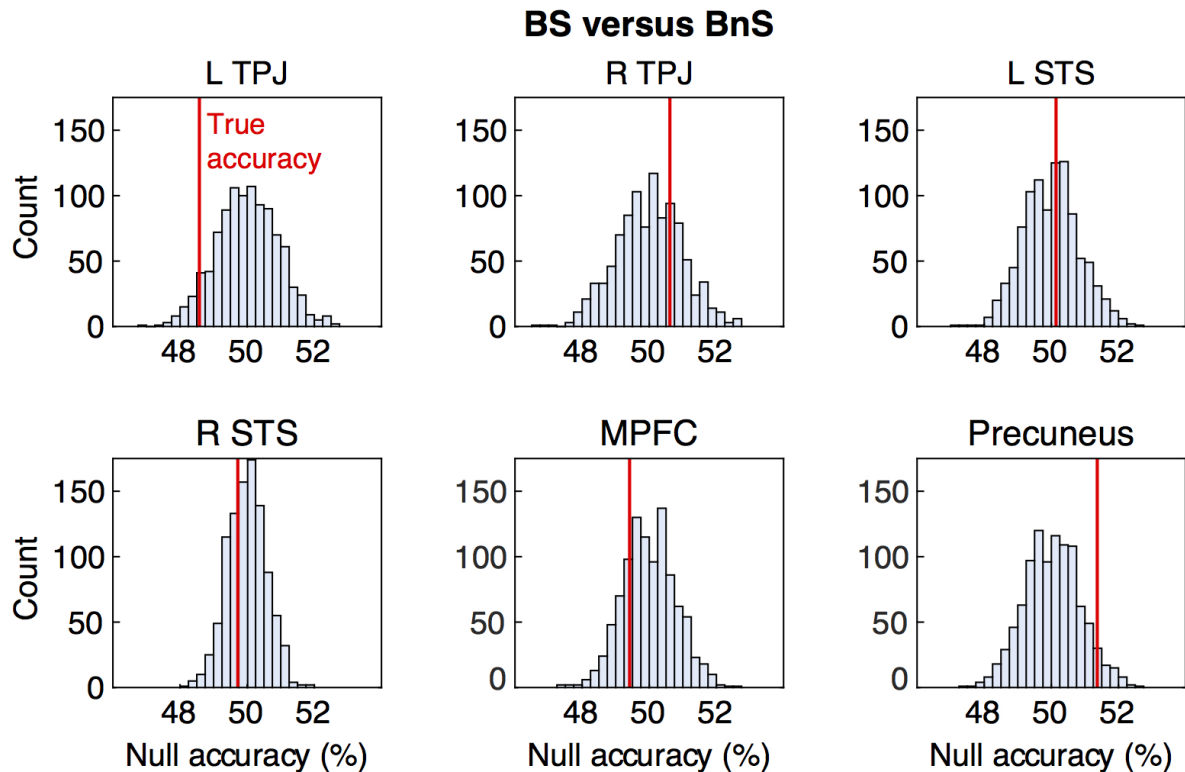


Figure 17. Decoding false belief from true belief trials. False belief trials were represented by the BS condition and matching true belief trials by the BnS condition. Each panel shows the results for one ROI. In each panel, the histogram shows the null distribution of decoding accuracies based on permutation testing with shuffled conditions labels (chance level = 50%). The red vertical line shows the accuracy of the classifier when it was trained and tested using the real (unshuffled) conditions labels. Significance threshold ($p < 0.05$) based on permutation testing, corrected for multiple comparisons across six ROIs using FDR. None of the ROIs showed significant decoding that distinguished false from true belief trials.

	L TPJ	R TPJ	L STS	R STS	MPFC	Precuneus
Decoding accuracy	48.60%	50.60%	50.10%	49.70%	49.40%	51.40%
95% CI	47.5 - 50.3	49.0 - 52.3	48.4 - 51.6	47.9 - 51.8	47.3 - 51.3	49.8 - 52.9
P value	0.936	0.26	0.42	0.683	0.776	0.053

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

ROI analysis, Hypothesis 2

Figure 18 shows the results for the ROI analysis in which nBS trials were compared with nBnS trials (see table 3 for test statistics). According to hypothesis 2, at least some of the ROIs should show significant decoding for this comparison, indicating sensitivity to switched versus non-switched trials in the specific case in which the head's vision is unblocked, by implication able to “see” whether the ball is switched or not. The results show a significant decoding for nBS versus nBnS trials in the right TPJ (the decoding accuracy was significantly above the chance level of 50%). This decoding result was significant even when corrected for multiple comparisons across the six defined ROIs ($p = 0.004$ uncorrected; $p = 0.024$ corrected). Moreover, the decoding in the right TPJ for nBS versus nBnS trials was significantly greater than the decoding in the right TPJ for BS versus BnS trials (52.8% versus 50.6%; $p = 0.0432$, permutation testing with 10,000 iterations). Thus, hypothesis 2 was supported by the present data. When subjects thought that the head could see the switch take place, then the right TPJ was affected by the switch (significant decoding for nBS versus nBnS); when subjects thought that the head could not see the switch take place, then none of the ToM ROIs were significantly affected by the switch (no significant decoding for BS versus BnS).

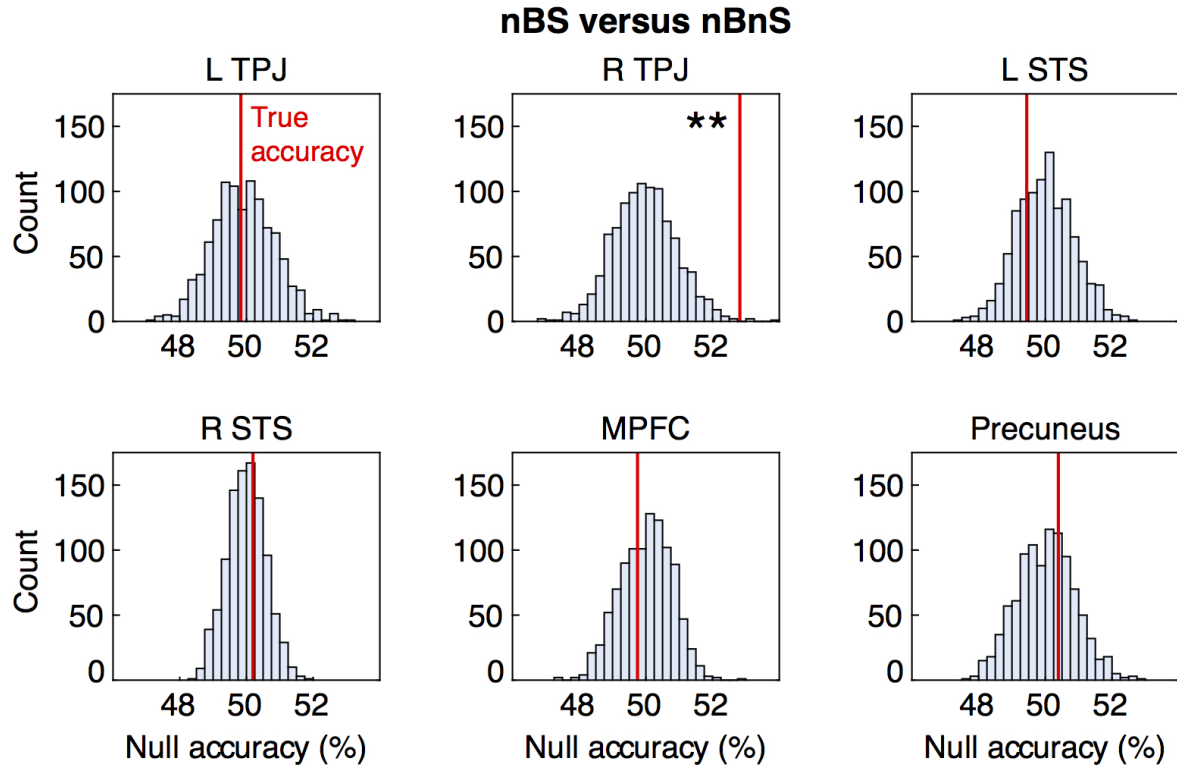


Figure 18. Decoding trials in which the cartoon “saw” a switch occur from trials in which the cartoon “saw” that no switch occurred. Trials in which the cartoon “saw” a switch were represented by the nBS condition, and trials in which the cartoon “saw” no switch were represented by the nBnS condition. Each panel shows the results for one ROI. In each panel, the histogram shows the null distribution of decoding accuracies based on permutation testing with shuffled conditions labels (chance level = 50%). The red vertical line shows the accuracy of the classifier when it was trained and tested using the real (unshuffled) conditions labels. A decoding accuracy significantly greater than chance is indicated by * ($p < 0.05$), based on permutation testing, corrected for multiple comparisons across six ROIs using FDR. The right TPJ showed significant decoding (p corrected = 0.024).

	L TPJ	R TPJ	L STS	R STS	MPFC	Precuneus
Decoding accuracy	49.80%	52.80%	49.40%	50.20%	49.70%	50.40%
95% CI	47.7 - 52.1	50.9 - 54.6	47.7 - 50.9	48.2 - 52.0	48.4 - 51.6	48.6 - 52.1
P value	0.518	0.004*	0.725	0.354	0.631	0.329

Table 3. Decoding trials in which the cartoon “saw” a switch from trials in which the cartoon “saw” no switch (nBS versus nBnS). Mean decoding accuracy (%), 95% confidence interval (based on bootstrap distribution), and p value (based on permutation testing, uncorrected for multiple comparisons) are shown

for each of the six ROIs. The * indicates significant p values that survived correction for multiple comparisons across all six ROIs (FDR-corrected $p < 0.05$).

Searchlight analysis

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

Post-hoc analyses

In addition to the tests for the two hypotheses described above, six additional exploratory analyses were run for the sake of thoroughly examining the data. These contrasts include the two main factors: all blocked trials versus all unblocked trials, and all switched trials versus all unswitched trials. The contrasts also include the remaining possible pairings among the four specific conditions: BS versus nBS, BS versus nBnS, BnS versus nBS, and BnS versus nBnS. For each contrast, we performed the MVPA analysis on the six defined ROIs. Figure 19 and Table 4 show the results. Only one of the additional post-hoc analyses showed significant decoding. The blocked versus unblocked contrast resulted in significant decoding in the precuneus (p uncorrected = 0.006; p corrected for 6 ROI = 0.036).

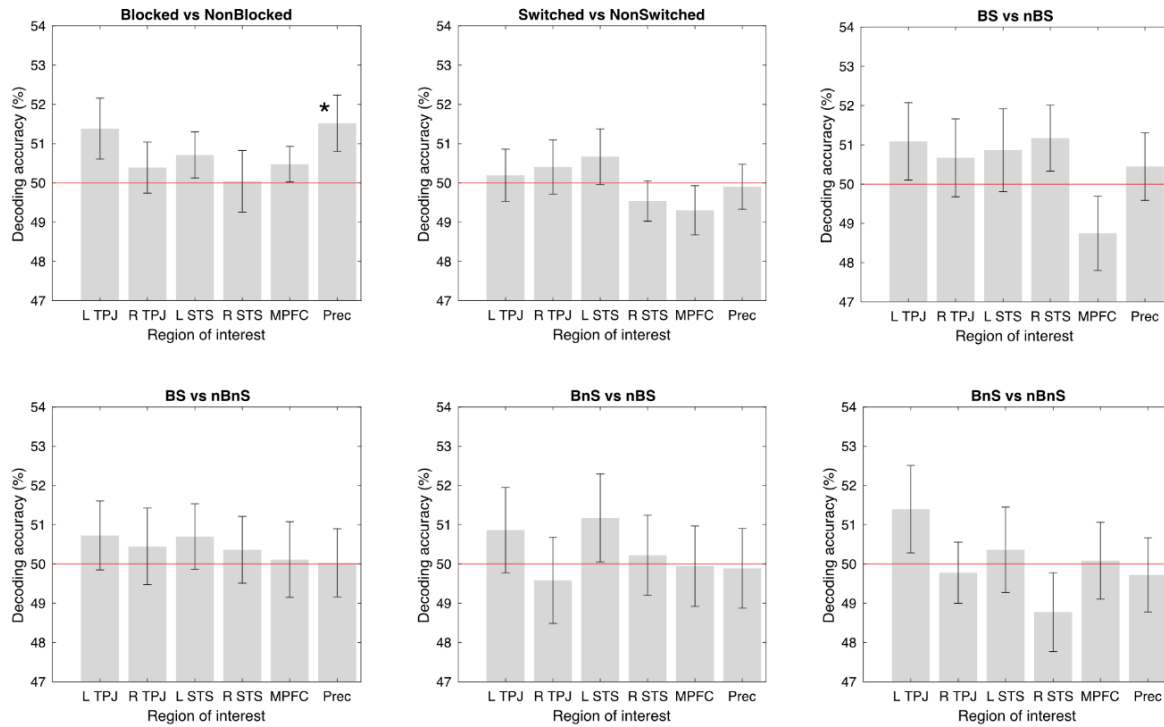


Figure 19. Six different contrasts, each performed within the six ROIs. Each bar shows mean decoding accuracy. Error bars show SEM. Red horizontal line indicates chance level decoding. Significance indicated by * ($p < 0.05$) based on permutation testing, corrected for multiple comparisons across the six ROIs using FDR.

		L TPJ	R TPJ	L STS	R STS	MPFC	Precuneus
Blocked versus Unblocked	Decoding accuracy (%)	51.4	50.4	50.7	50	50.5	51.5
	95% CI	49.7 to 52.8	49.1 to 51.6	49.7 to 51.9	48.5 to 51.4	49.6 to 51.4	50.2 to 53.2
	P value	0.021	0.269	0.115	0.448	0.173	0.006*
Switched versus Unswitched	Decoding accuracy (%)	50.2	50.4	50.7	49.5	49.3	49.9
	95% CI	48.9 to 51.6	49.0 to 51.6	49.2 to 52.0	48.6 to 50.6	48.1 to 50.5	48.8 to 51.1
	P value	0.373	0.275	0.126	0.88	0.912	0.588
BS versus nBS	Decoding accuracy (%)	51.1	50.7	50.9	51.2	48.7	50.4
	95% CI	49.3 to 53.1	48.7 to 52.5	48.9 to 53.0	49.4 to 52.7	46.9 to 50.6	48.6 to 51.9

	P value	0.137	0.243	0.163	0.021	0.96	0.305
BS versus nBnS	Decoding accuracy (%)	50.7	50.4	50.7	50.4	50.1	50
	95% CI	49.2 to 52.5	48.6 to 52.3	49.1 to 52.4	49.0 to 52.1	48.3 to 52.0	48.4 to 51.6
	P value	0.218	0.318	0.199	0.252	0.409	0.462
BnS versus nBS	Decoding accuracy (%)	50.9	49.6	51.2	50.2	49.9	49.9
	95% CI	48.7 to 53.0	47.3 to 51.5	49.1 to 53.2	48.2 to 52.1	48.0 to 52.2	47.9 to 51.9
	P value	0.183	0.678	0.081	0.32	0.515	0.529
BnS versus nBnS	Decoding accuracy (%)	51.4	49.8	50.4	48.8	50.1	49.7
	95% CI	49.1 to 53.5	48.4 to 51.4	48.2 to 52.7	46.7 to 50.6	48.2 to 51.8	48.0 to 51.7
	P value	0.071	0.583	0.329	0.976	0.459	0.633

Table 4. Six different contrasts, each performed within the six ROIs. Mean decoding accuracy (%), 95% confidence interval (based on bootstrap distribution), and p value (based on permutation testing, uncorrected for multiple comparisons) are shown for each of the six ROIs. The * indicates significant p values that survived correction for multiple comparisons across all six ROIs (FDR-corrected $p < 0.05$).

Discussion

The present experiment used fMRI to measure brain activity during a visual, reaction-time, ToM task that incorporated both false-belief and true-belief trials. We used the task to test two specific hypotheses. In hypothesis 1, brain areas in a previously described ToM cortical network should respond differently to false-belief trials than to true-belief trials (contrasting BS and BnS trials). This hypothesis was not supported by the data. We found no significant decoding of false-belief versus true-belief trials in any of the six ROIs that we tested. Hypothesis 2 was offered as an alternative. In hypothesis 2, the ToM cortical network should not respond differently to BS and BnS trials, because the cartoon head had its vision blocked in both conditions – subjects knew that it could not “see” the ball switch – and thus, by implication, the mental state of the head should have been unaffected by whether the ball switched or not. The

ToM brain areas, building a model of the mental state of the agent, therefore should not react differently in those two conditions. Instead, only when the head was unblocked, and by implication could see whether the ball switched or not, should the ToM network react differently to the switch and non-switch conditions. Hypothesis 2 therefore predicted significant decoding in the nBS versus nBnS comparison. This hypothesis was supported by the data. The right TPJ showed decoding significantly above chance for the nBS versus nBnS comparison.

A possible alternative explanation of the present results is that the right TPJ reacted to the difference between switched and unswitched trials because of a processing cost, or other general effect on processing, caused by the subjects seeing the ball switch from one box to another. The average percent correct was indeed slightly lower (by an average of 2.15%), and the average latency higher (by 33.49 ms or 3.4%), in switched conditions than in unswitched conditions (see Results, Task performance), suggesting a processing cost. However, this explanation of the TPJ result is unlikely. For BS versus BnS trials, we found no significant signal in the right TPJ; for nBS versus nBnS trials, we did find a significant signal in the right TPJ. If the right TPJ activity reflected a difference between switched and unswitched trials, it was evidently not a general effect, but only occurred when the subjects thought that the cartoon character could see the switch take place. We suggest, therefore, that our interpretation in terms of modeling the mind states of others is the most likely.

The present results might also help to explain the somewhat mixed results of previous studies that compared false belief and true belief conditions (Aichhorn et al., 2009; Doehnel et al., 2012; Hooker et al., 2008; Sommer et al., 2010). On the one hand, false belief conditions may require more cognitive complexity on the part of the subject to infer another agent's mental state. For that reason, one might hypothesize that the ToM cortical network should be more

active in false belief trials than in true belief trials. On the other hand, the implied mental state of the agent in question is not necessarily different in false versus true belief trials. Thus, modeling the same mental state, the ToM cortical network might respond in the same way to false and true belief trials. The distinction between false and true belief, therefore, may be a less incisive test of the ToM network than the distinction between different mental states attributed to an agent.

On the basis of the current data, we cannot rule out the possibility that brain areas in the ToM network, such as the right TPJ, react specifically to the distinction between false and true beliefs in others. One important difference that may have contributed to previous work finding a difference where we did not could be the use of univariate analysis compared with our MVPA approach. A univariate analysis involves a contrast between the magnitude of BOLD signal for each condition, so if each condition evokes the same magnitude signal in the same region(s), even if they are unique patterns of activity, no distinction will be found. The MVPA approach we employed instead compares patterns of activity in regions rather than overall magnitude differences. Because our paradigm was intentionally so similar between conditions, we did not expect univariate analyses to detect differences between conditions, so MVPA was used instead.

Another potential factor that could explain why this study did not detect a difference in the false vs true-belief conditions is that our design incrementally introduced information during the trial. At least one study (Aichhorn et al., 2009) suggests that false beliefs are processed at the moment they are first detected, and not at a later time when a question is asked about them. For the sake of caution, therefore, what we can infer from our results is that in the particular task that we used, when analyzing brain activity at the moment of a reaction-time decision, the ToM network appears to place greater emphasis on reconstructing the internal mental experience of another agent than it does on the outside perspective or knowledge that allows the subject to

evaluate whether that agent has a false or true belief. Compared to the findings from Chapters 3 & 4 these results may be surprising. The false-belief condition seemed uniquely important when only behavioral measures were of interest, i.e., reaction time for the study in Chapter 3 and accuracy for the study in Chapter 4, however when neural activity was measured, the false-belief condition could not be dissociated from a true-belief condition. This apparent conflict may be explained by the fact that false beliefs are useful in determining ToM abilities as they ensure the subject is thinking about another agent, but that doesn't necessarily equate to differences in neural signatures. Chapters 3 and 4 are in line with previous ToM literature showing how false beliefs can be useful for measuring how people think about others. This chapter, although it challenges some previous findings, demonstrates that individuals represent rich details of other's mental experiences.

Chapter 6 - Modeling types of attention in the self and others

Introduction

Reconstructing someone else's attentional state is of central importance in theory of mind (Baron Cohen, 1997; Calder et al., 2002; Graziano, 2013). By identifying the object of someone else's attention and having some intuitive understanding of the complex dynamics and consequences of attention, one can reconstruct at least some of the other person's likely thoughts, intentions, and emotions, and make predictions about that person's behavior. As was previously described in Chapter 2, to date, the majority of work on how people reconstruct the attention of others has focused on gaze direction. Kobayashi and Kohshima (1997) suggest the high contrast between the pupil and sclera in the human eye as a possible adaptation for better gaze tracking. Other work shows that seeing a face gaze at an object automatically draws one's own attention to the object (Friesen & Kingstone, 1998; Frischen et al., 2007). These and other findings show the importance of reconstructing gaze direction in social cognition.

To be adaptive in aiding theory of mind, however, a model of attention should be far more than a vector indicating gaze direction. Graziano and colleagues have previously suggested that the human brain constructs a rich, dynamic, and predictive model of other people's attention (Graziano, 2019; Graziano, 2013; Graziano & Kastner, 2011; Wilterson et al., 2021). This model, if successful should contain information about different types of attention, about the temporal dynamics with which attention tends to move from item to item, about how external factors such as salience and distractions are likely to affect a person's attention, and about how attention interacts with other cognition, memory, and behavior. In the proposal, that model is informed and constrained by incoming information, including gaze direction; however, gaze is not the sole factor. In the real world, people rely on a host of factors such as facial expression,

body language, speech, and social context. For example, blind people must be able to build models of other people's attention without seeing the eyes of those they interact with. Likewise, during a phone conversation, we cannot see the person we are communicating with and yet we can often know when they are engaged with us or if they are distracted.

Recent work provides evidence for an automatically constructed model of the attention of others that may go beyond simply registering gaze direction (Kelly et al., 2014; Pesquita et al., 2016; Vernet et al., 2019). Pesquita and colleagues (2016) found that when participants watch an actor in a video attend to an object, the participants implicitly distinguish between whether the actor's attention is drawn to the object exogenously (bottom-up or stimulus-driven attention), or whether the actor endogenously shifts attention to the object (top-down or internally driven attention). Exogenous and endogenous attention are the two principal ways in which selective attention moves between objects. They are emphasized in distinct cortical networks (the ventral and dorsal attention networks), and they influence the behavior of agents in profoundly different manners (Corbetta et al., 2008; Corbetta & Shulman, 2002; Posner, 1980; Shulman et al., 2010). The ability to identify, track and distinguish between someone else's exogenous and endogenous attention is therefore one example of how people may construct a rich, dynamic, and useful model of other people's attention beyond merely encoding gaze direction or identifying the object of attention.

In the present study, functional magnetic resonance imaging and multi-voxel pattern analysis were used to study brain activity in participants while they read brief stories about people's attention (Figure 20A). The task consisted of stories involving an agent whose attention was being attracted exogenously ('Kevin walks into his closet and notices the bright red tie. . .') and stories involving an agent whose attention was being directed endogenously ('Kevin walks

into his closet and looks for the bright red tie. . .'). Analogous first-person stories were also included which cast the subject of the experiment as the agent ('You walk into your closet and notice the bright red tie. . .'). The study therefore used a 2X2 design (Figure 20B; exogenous versus endogenous attention and 'self' agent versus 'other' agent). Finally, a fifth, control condition was included which consisted of nonsocial stories in which the agent was replaced by an inanimate object, that, like attention, has a source and a target, such as a camera or a light source ('In a closet, a light shines on a red tie. . .').

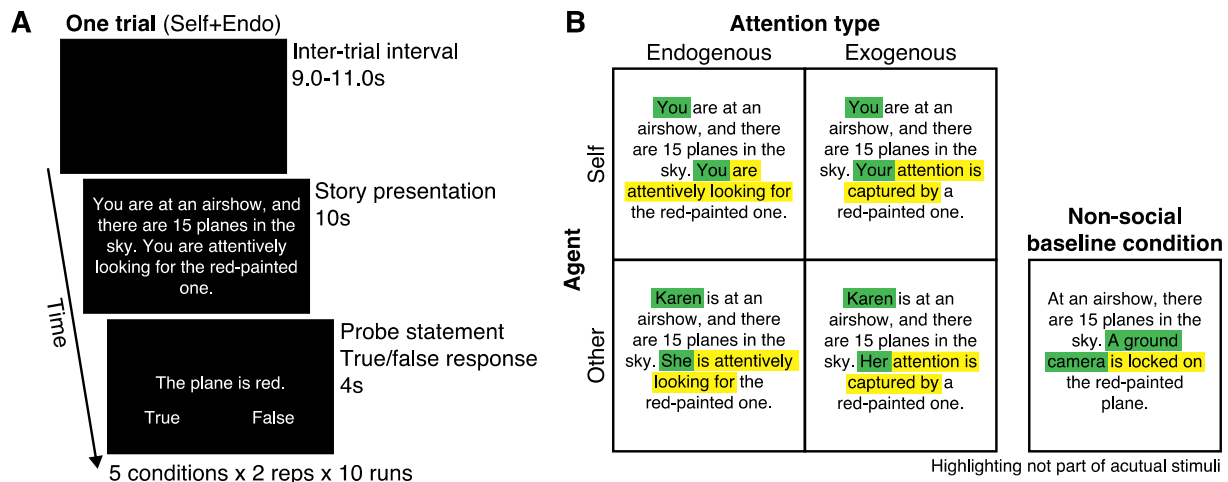


Figure 20. (A) Schematic timeline of the fMRI design. In each trial, subjects were presented with a short story for 10 s, describing a scene in which an agent attended to an object in the environment. A probe statement was then shown for 4 s, relating to either the story's spatial context or object property, to which the subjects responded either true or false by button press. (B) The agent in the story was either the subject him-/herself (self) or another person (other) and directed attention to the object endogenously or exogenously. Green highlighting indicates wording specifying agent, yellow highlighting indicates wording specifying attention type (colors not part of actual visual stimuli). A nonsocial control condition was also included, in which the agent was replaced by a non-human object.

This design allowed for four predictions which were inspired by the AST as well as the fMRI study described in Chapter 5. AST predicts that individuals create rich models of attention for themselves and for others, and the results of Chapter 5 also demonstrate that we create rich models of others' awareness. Building from those premises, we hypothesized that participants

would encode the type of attention in the story (exogenous versus endogenous), and that this encoding would be evident in areas classically involved in theory of mind. This first prediction, that the social cognition network will encode the exogenous-versus-endogenous distinction, represents the main, novel contribution of this study. Previous studies have used MVPA to decode various aspects of other people's mental states from activity in social brain areas, such as their beliefs (Koster-Hale et al., 2017), intentions (Koster-Hale et al., 2013), and perceptual source (Koster-Hale et al., 2014). To the best of our knowledge, this investigation is the first to test whether activity in social brain areas can decode other people's attentional states.

The second prediction was that participants would encode information about the agent in the story (self versus other), and that this encoding would again be evident in regions of the theory-of-mind network. Self-versus-other encoding has been examined in previous studies and found to be reflected in these areas (e.g., Northoff et al., 2006; Ochsner et al., 2004; Passingham et al., 2010; Qin & Northoff, 2011; van Veluw & Chance, 2014). This second prediction represents a test of whether the present paradigm, using subtle wording differences between similar sentences, can produce results consistent with previous findings.

Third, it was predicted that participants would encode information associated with the interaction between the two factors. We predicted that areas in the theory-of-mind network may encode the type of attention (exogenous versus endogenous) to a different extent in self-related stories as compared to other-related stories.

Fourth and finally, we tested for brain regions that encoded the distinction between social stories (with human agents) and nonsocial stories (with only non-agent objects). This social-versus-nonsocial encoding was predicted to also be evident in the same network of brain regions noted above, that are known to be involved in theory of mind. This final analysis served as a

control to check on the validity of the story stimuli and confirm that they engaged social cognition as expected.

This chapter describes an fMRI experiment on human subjects, which was published as *Temporo-parietal cortex involved in modeling one's own and others' attention* (Guterstam et al., 2021). My collaborator, Dr. Guterstam was the primary author on the experiment presented in this chapter, but it is included here because I contributed to the development of the task and because this study provides valuable contextual information which aids in the understanding of the other work contained in this dissertation, for which I was the principal experimenter.

Methods

Subjects

Thirty-two subjects (12 female, 30 right-handed, aged 18–52, normal or corrected to normal vision) participated in the study. Subjects were recruited either from a paid subject pool, receiving 40 USD for participation, or from among Princeton undergraduate students, who received course credits as compensation. In the subject recruitment material, the experiment was described as a ‘Reading Comprehension Study.’ All subjects provided informed consent and all procedures were approved by the Princeton Institutional Review Board.

Experimental setup

Before scanning, subjects were given instructions and then shown three sample trials (which did not belong to the set of stories presented in the experiment) on a laptop computer screen. All subjects gave the correct responses to all three trials on the first try, indicating they had understood the instructions. While scanning, subjects laid in a supine position on the MRI bed. Through an angled mirror mounted on top of the head coil, they viewed a screen

approximately 80 cm from the eyes, on which visual stimuli were projected with a Hyperion MRI Digital Projection System (Psychology Software Tools, Sharpsburg, PA, USA) with a resolution of 1920 x 1080 pixels. A PC running MATLAB (MathWorks, Natick, MA, USA) and the Psychophysics Toolbox (Brainard, 1997) was used to present visual stimuli. A right-hand 5-button response unit (Psychology Software Tools Celeritas, Sharpsburg, PA, USA) was strapped to subjects' right wrist. Subjects used the right index finger button to indicate a true response, and the right middle finger to indicate a false response during the probe phase of each trial.

Experimental conditions and stimuli

The experiment consisted of five experimental conditions. Subjects were presented with short stories (2–3 sentences, average word count = 24) describing a scene in which an agent, either the subject him-/herself (self) or another person (other), directed attention to something in the external world endogenously (e.g., “X is attentively looking for Y”) or exogenously (e.g., “X’s attention is captured by Y”). These four conditions made up a 2X2 factorial design: attention type (endogenous versus exogenous) by agent type (self versus other). In addition, a fifth, control condition was included that featured stories in which the agent was substituted by a non-human object. In each trial, after a 9–11 s inter-trial interval, the story was presented for 10 s in easily readable, white text on a black background, at the center of the screen. After each story a probe statement was shown for 4 s, which required subjects to respond either true or false by button press.

Each subject performed 100 trials which consisted of 80 social stories and 20 nonsocial control stories. The 80 social stories were constructed as follows. We began with 80 unique short stories. For each story, four versions were constructed, one for each of the factorial conditions (Figure 20B). To keep the story versions as semantically similar as possible, we made minimal

changes to the wording between conditions. To distinguish the self and other versions, we substituted the word ‘you’ with a name (e.g., ‘Andrew’) and the word ‘your’ with ‘his’ or ‘her’. The names in the stories were selected from a list of the 100 most popular given names for male and female babies born during the years 1919–2018 in the United States, which is published by the Social Security Administration. Half of the names were masculine, half feminine. To distinguish the endogenous and exogenous story versions, we used different wording for the part of the story where the agent (X) is related to the object (Y). In the endogenous versions, we used formulations such as: “X is trying to find Y,” “X is trying to spot Y,” or “X is looking attentively for Y.” The intent was to capture the intentional nature of the agent to willfully engage in an action that was self-initiated. In the exogenous versions, we used formulations such as: “X’s eyes are drawn to Y,” “X’s gaze is captured by Y,” or “X’s attention is captured by Y.” Here, intending to express the pull of the agent’s attention toward an object by outside forces. We matched the average number of words across all four conditions (24 words). The number of stories that included the words ‘attention’ or ‘attentively’ was balanced between the endogenous and exogenous categories (43 stories in each). Among the 80 stories, for each subject, 20 were randomly selected to be used in the endogenous-self version; 20 in the endogenous-other version; 20 in the exogenous-self version; 20 in the exogenous-other version. Thus, for the example story shown in Figure 20B, each subject saw only one of the four versions. Therefore, the 100 unique stories were balanced for as many properties as possible other than the two factors that were manipulated.

Finally, we constructed 20 additional stories for the nonsocial control condition (Figure 20B). To keep the control stories as semantically similar as possible to the social stories, we based them on a subset of the 80 original stories. Crucially, the agent in the original story was

substituted with a nonhuman object, such as a camera or a spotlight, that has a source and a target just as attention does. For instance, the original story, "You are in a bike shop, and numerous bikes hang on one of the walls. You are attentively looking for that red Italian sports bike", was adapted to the nonsocial condition by substituting the agent with a spotlight: "In a bike shop, on one of the walls, hangs numerous bikes. A bright spotlight is shining on a red Italian sports bike". The average number of words of the nonsocial stories (24 words, standard deviation = 3) was matched with the attention stories.

The purpose of the probe statement at the end of each trial was to ensure that subjects carefully read the stories. Each statement described one detail of the preceding story that could either be true or false. We restricted the probe statements to a location detail of the story (place probe: e.g., "Emma is on a bus") or the description of the object (object probe: e.g., "The Van Gogh painting has sunflowers") in order to avoid alerting subjects to the focus of the experiment on theory of mind and attention. Half of the probe statements were place probes and half object probes. Within both the place and the object probes, half were true, and half were false. The probe was on screen for 4 s, during which subjects were required to indicate whether the statement was true or false by button press.

The experiment consisted of 10 runs of approximately 4 minutes each. In each run, the five conditions were repeated two times, yielding a total of 10 trials per run. The trial order was randomized, with the limitation that two consecutive trials could not belong to the same condition. Each run included 18 s of baseline before the onset of the first trial and 12 s of baseline after the conclusion of the last trial.

fMRI data acquisition

Functional imaging data were collected using a Siemens Prisma 3T scanner equipped with a 64-channel head coil. Gradient-echo T2*-weighted echo-planar images (EPI) with blood-oxygen dependent (BOLD) contrast were used as an index of brain activity (Logothetis et al., 2001). Functional image volumes were composed of 54 near-axial slices with a thickness of 2.5 mm (with no interslice gap), which ensured that the entire brain excluding cerebellum was within the field-of-view in all subjects (54 x 78 matrix, 2.5 mm x 2.5 mm in-plane resolution, TE = 30 ms, flip angle = 80°). Simultaneous multi-slice (SMS) imaging was used (SMS factor = 2). One complete volume was collected every 2 s (TR = 2000 ms). A total of 1300 functional volumes were collected for each participant, divided into 10 runs (130 volumes per run). The first three volumes of each run were discarded to account for non-steady-state magnetization. A high-resolution structural image was acquired for each participant at the end of the experiment (3D MPRAGE sequence, voxel size = 1 mm isotropic, FOV = 256 mm, 176 slices, TR = 2300 ms, TE = 2.96 ms, TI = 1000 ms, flip angle = 9°, iPAT GRAPPA = 2). At the end of each scanning session, matching spin echo EPI pairs (anterior-to-posterior and posterior-to-anterior) were acquired for blip-up/blip-down field map correction.

fMRI preprocessing

The same preprocessing procedures described in Chapter 5 were employed for the current study.

Testing prediction 1

The purpose of the first analysis was to determine whether the type of attention (endogenous or exogenous) present in the stories could be uniquely represented in the brain. For

this analysis, we used MVPA, due to its more sensitive capability compared to standard univariate subtraction methods. The reason for using this sensitive analysis is that the difference between exogenous and endogenous trial types was extremely subtle. Both trial types engaged social cognition, and therefore might cancel each other out in a simple subtraction. The stimuli were nearly identical, differing only in a few words that indicated the type of attention used by the agent in the story. In addition, the type of attention featured in the story was irrelevant to the task performed by the subject. To accommodate the subtlety of the distinction between conditions, the study was designed to use MVPA. It was hypothesized that brain patterns would carry information about the endogenous versus exogenous distinction; and that decoding would be successful in ROIs within the network of areas typically found to be involved in social cognition.

For a complete description of the MVPA procedures used for this study, see Chapter 5 under the MVPA Analysis section. In addition to the ROI analysis, a further exploratory whole-brain searchlight analysis (Kriegeskorte et al., 2006) was used to test for possible areas of decoding outside the ROIs.

Testing prediction 2

The purpose of the second analysis was to determine whether the brain encoded information concerning the type of agent (self versus other) present in the stories. The analysis methods were the same as for testing hypothesis 1, except that the self-related and other-related trials regressors of interest were used, collapsed across the type of attention (exogenous or endogenous). Just as for hypothesis 1, analyses were conducted in the six defined ROIs within the theory-of-mind network.

Testing prediction 3

The purpose of the third analysis was to test for an interaction between the two variables of interest (attention type and agent type). MVPA was again used to test whether the decoding for the type of attention was significantly different between the self-related and the other-related stories. The analysis methods were similar to those used for testing hypotheses 1 and 2, except in the following ways. Two MVPA decoding results were computed, the first for distinguishing endogenous-self from exogenous-self stories, the second for distinguishing endogenous-other from exogenous-other stories. We then computed the difference between the two decoding results ([endogenous-self versus exogenous-self] – [endogenous-other versus exogenous-other]) to create a decoding difference score. Just as for hypotheses 1 and 2, we tested the six defined ROIs within the theory-of-mind network (van Veluw & Chance, 2014).

In a post hoc analysis, a two-way cross-classification analysis was used to test for overlap in attention type encoding in self and others in the left TPJ ROI. One classifier was trained to discriminate endogenous versus exogenous self-stories and tested on other-stories (using a leave-one-runout approach), and another classifier was trained to discriminate endogenous versus exogenous other-stories and tested on self-stories, from which an average cross-classification decoding accuracy was obtained. Overlap in attention type representations in self and other should be reflected in above-chance decoding.

Testing prediction 4

The purpose of the fourth analysis was to confirm whether the story stimuli engaged social cognition and thereby recruited brain areas within the expected theory-of-mind network. The analysis was meant as an added control to check the validity of the paradigm. The analysis

methods were similar to those used for testing hypotheses 1–3, except in the following ways. We computed four MVPA decoding results: endogenous-self versus nonsocial, endogenous-other versus nonsocial, exogenous-self versus nonsocial, and exogenous-other versus nonsocial. It was not possible to compare all 80 social trials to the 20 nonsocial trials, because MVPA requires equal numbers of trials in both conditions for comparisons. Each analysis represents a separate, alternative way to assess the social-versus-nonsocial decoding. Just as for hypotheses 1–3, we tested the six defined ROIs within the theory-of-mind network.

Results

Behavioral results

Participants answered a simple true-or-false probe question after each story (e.g., “Emma is on a bus”), to ensure alertness throughout the experiment. Although the MRI results depended on the time interval during the reading of the story and not during the reading or answering of the probe question, the behavioral response to the probe question may give some indication of whether the story conditions were well balanced. If one type of story was more difficult to process or caused the subjects to think more deeply about the character in the story, that difference may be reflected in a different accuracy or latency in responding to the subsequent probe question. However, no significant differences were observed in accuracy ($F = 0.15$, $p = 0.930$, ANOVA) or latency ($F = 1.66$, $p = 0.181$, ANOVA) among the four social story conditions.

When participants responded to the probe questions after nonsocial, control stories, versus when they responded to the probe questions after social stories, no significant difference in accuracy was found ($t = 1.83$, $p = 0.077$), although as expected, a slightly longer latency was

observed after nonsocial stories than after social stories ($t = 9.18$, $p < 0.001$). The reason for the latency difference is likely because the probe statements in the nonsocial condition were on average two words longer than those in the social conditions (eight words versus six words), because the character in the probe statements ("You . . ." or "Emma . . .") was replaced by an object that required more words to describe (e.g., "The surveillance camera . . ."). One might therefore expect that it took participants a little longer to read the nonsocial probe statements compared to the social ones. We suggest that this subtle difference in latency during the post-story question period is unlikely to have affected the comparison of MRI activity between social and nonsocial conditions, since the relevant MRI activity was evoked by the time period during the reading of the story, not during the reading and answering of the questions.

Prediction 1

We hypothesized that participants would encode the attentional state of the agents in the stories in enough detail to distinguish between endogenous and exogenous attention, even though the difference between the story types was extremely subtle – only a few words that vary slightly altered the semantic meaning of the sentences. We made the strong prediction that decoding would be found within the set of brain areas typically included in the theory-of-mind cortical network. The same set of theory-of-mind ROIs described in Chapter 5 were used for this study (see Figure 16). Figure 21A shows the results and the statistics can be found in table 5. Decoding accuracy for endogenous versus exogenous stories was significantly above chance for the left TPJ, and the significance of the left TPJ decoding survived a multiple comparison correction for the six ROIs (mean decoding accuracy 52.9%, 95% CI 50.7–55.2, p -uncorrected = 0.0046, p FDR-corrected = 0.0276). The results therefore show that activity in the left TPJ allowed for

significant decoding of the attentional state – exogenous versus endogenous – of agents in a story.

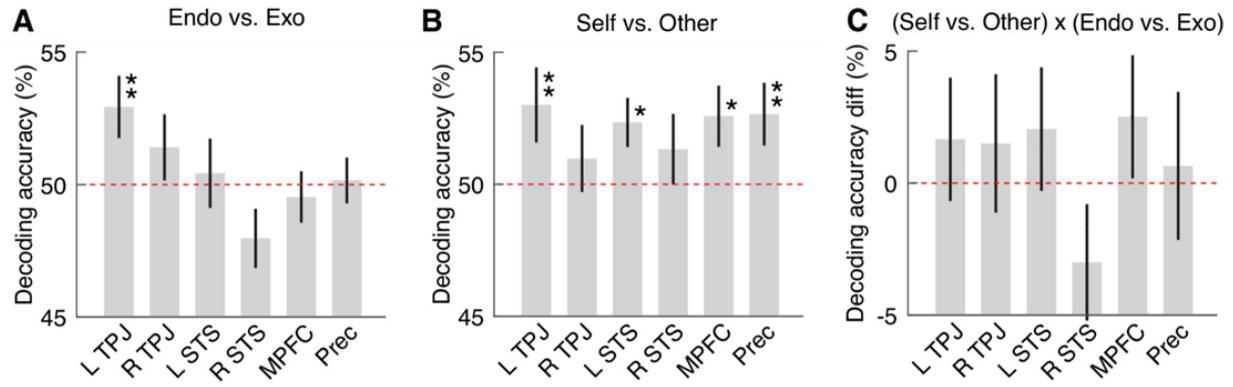


Figure 21. Decoding attention type, agent, and the interaction between them, in six brain areas. Each point shows mean decoding accuracy. Error bars show SEM. Red horizontal line indicates chance level decoding. Significance indicated by * ($p < 0.05$) and ** ($p < 0.01$), based on permutation testing (all significant p values also survived FDR correction for multiple comparisons across all six ROIs [all significant corrected p s < 0.05]). (A) The ability of a classifier, trained on BOLD activity patterns within each ROI, to decode endogenous (endo) versus exogenous (exo) attention. (B) Decoding accuracy for agent (self versus other). (C) Decoding accuracy for the interaction between type of attention and agent.

		L TPJ	R TPJ	L STS	R STS	MPFC	Precuneus
Endo vs. Exo	Mean accuracy	52.90%	51.40%	50.40%	48.00%	49.50%	50.20%
	95% CI	50.7–55.2	49.1–53.9	47.8–52.8	45.9–50.1	47.6–51.4	48.5–51.8
	P value	0.0046*	0.1148	0.3518	0.9547	0.6439	0.4428
Self vs. Other	Mean accuracy	53.00%	51.00%	52.30%	51.30%	52.60%	52.70%
	95% CI	50.1–55.6	48.5–53.4	50.6–54.1	48.9–54.1	50.5–55.0	50.4–55.0
	P value	0.0053*	0.1974	0.0204*	0.1241	0.0105*	0.0099*
(Self vs. Other) × (Endo vs. Exo)	Mean accuracy diff	1.60%	1.50%	2.00%	–3.0%	2.50%	0.60%
	95% CI	–2.7–6.3	–3.6–6.5	–2.7–6.3	–7.4–1.1	–2.3–6.7	–5.5–5.5
	P value	0.243	0.2639	0.1967	0.8944	0.1414	0.39

Table 5. Decoding attention type, agent, and the interaction between the two, within the six ROIs. Mean decoding accuracy (%), 95% confidence interval (based on bootstrap distribution), and p value (based on permutation testing) are shown for each of the six ROIs. Results shown for decoding endogenous (endo) versus exogenous (exo) attention type, self versus other agent type, and the interaction between the two variables. * indicates significant p values that survived correction for multiple comparisons across all six ROIs (FDR corrected $p < 0.05$).

A searchlight analysis was performed over the whole brain to test whether any further areas may have significantly decoded the endogenous-versus-exogenous distinction. We found no brain-wide significant clusters of decoding for the endogenous versus exogenous distinction.

To control for potential univariate effects that could drive classifier performance, the endogenous > exogenous and exogenous > endogenous univariate contrasts were explored, which did not reveal any significant activity. These findings are compatible with previous studies (e.g., Hassabis et al., 2009) that have demonstrated the superiority of pattern-sensitive multivariate analyses compared with conventional univariate approaches for detecting differences in activity between conditions with highly similar macroscopic characteristics.

In addition to these planned analyses, the endogenous-versus-exogenous decoding was also explored within dorsal attention network regions. This analysis was motivated by an alternative hypothesis: people might simulate the act of attention orienting when reading the exogenous and endogenous stories, and thus activate the corresponding ventral (exogenous) attention network, to which the TPJ belongs, and dorsal (endogenous) attention network in a ‘mirror-neuron-like’ fashion. However, this control analysis revealed no significant decoding in any of the dorsal attention network ROIs.

Prediction 2

Our second prediction was that participants would differently represent the two types of agents in the stories (self versus other). We made the strong prediction that decoding would be

found within the same set of ROIs in the theory-of-mind cortical network. Figure 21B shows the results and the statistics can be found in table 5. Decoding accuracy for self versus other stories was significantly above chance, and survived a multiple comparisons correction, for the left TPJ (mean decoding accuracy 53.0%, 95% CI 50.1 to –55.6, p -uncorrected = 0.0053, p FDR-corrected = 0.0210), left STS (mean decoding accuracy 52.3%, 95% CI 50.6–54.1, p -uncorrected = 0.0204, p FDR-corrected = 0.0306), MPFC (mean decoding accuracy 52.6%, 95% CI 50.5–55.0, p -uncorrected = 0.0105, p FDR-corrected = 0.0210), and precuneus (mean decoding accuracy 52.7%, 95% CI 50.4–55.0, p -uncorrected = 0.0099, p FDR-corrected = 0.0210). These results confirm that the present paradigm, using stories that are subtly different from each other, can obtain social cognitive results that are consistent with previous findings.

Prediction 3

We hypothesized that areas in the theory-of-mind network would not only encode the distinction between endogenous and exogenous attention but do so to a significantly different extent in self-related stories than in other-related stories. However, the results showed no significant interaction in any of the ROIs (Figure 21C and Table 5). Thus, we found no support for prediction 3.

An alternative hypothesis is that attention type is encoded similarly in self and others. In a post hoc analysis, we focused on the left TPJ which was the only ROI that showed significant attention type decoding (Prediction 1) and tested for overlap in attention encoding in self and others using a two-way cross-classification analysis. In this analysis, one classifier was trained to discriminate endogenous versus exogenous self-stories and tested on other-stories, and another classifier was trained to discriminate endogenous versus exogenous other-stories and tested on self-stories, from which an average cross-classification decoding accuracy for the left TPJ was

obtained. Endogenous-versus-exogenous decoding significantly generalized across self-stories and other-stories (mean decoding accuracy 51.9%, 95% CI 49.9–54.1, $p=0.0393$), suggesting at least some degree of overlap in the encoding of attention in others and in oneself.

Prediction 4

Finally, we asked whether the activity in the theory-of-mind network would distinguish between social stories and nonsocial stories. This final analysis served as a control to check the validity of the story stimuli and confirm that they engaged social cognition as expected. We expected a signal of much greater magnitude in this analysis than in the analyses described above. The reason is that, as noted above, the types of social stories differed from each other by only a few words, and were nearly identical in semantic content; thus, any brain signal reflecting those differences is expected to be subtle. The distinction between social and nonsocial stories, however, was much greater semantically, and therefore the evidence of decoding in the brain is expected to be of greater magnitude.

Figure 22 shows the results, and the statistics can be found in table 6. The results are separated into six ROIs, and for each ROI, separated into four individual analyses, corresponding to each of the four main social conditions contrasted with the nonsocial control. Decoding accuracy was significantly greater than chance in almost all analyses across the six ROIs. The right STS showed the least consistent evidence of decoding. The TPJ bilaterally and the precuneus showed the most consistent evidence of decoding. These results show strong evidence of decoding of the social versus nonsocial stimuli in the known theory-of-mind, cortical network.

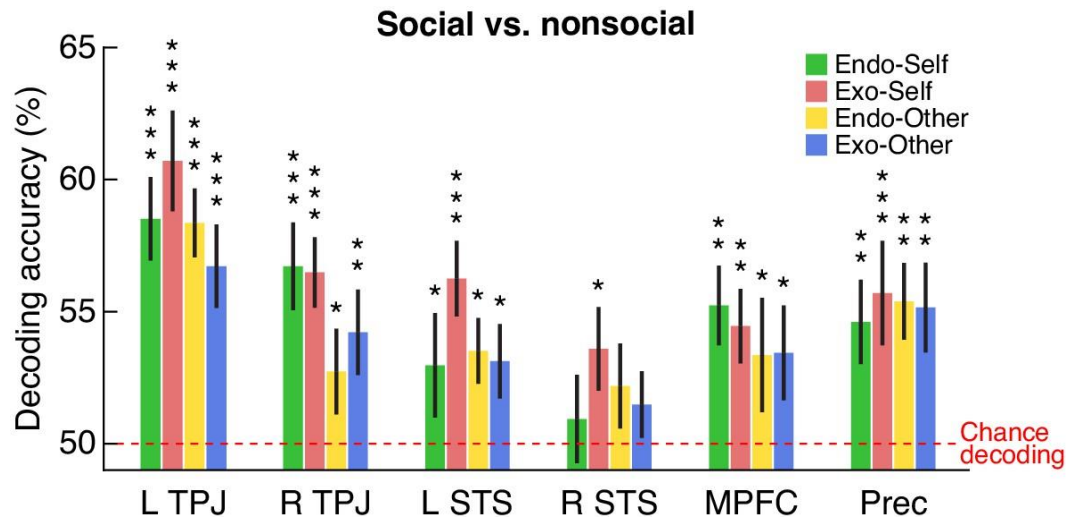


Figure 22. Decoding social versus nonsocial stories. The ability of a classifier, trained on BOLD activity patterns within each of the six ROIs, to decode each of the four social story conditions (endogenous-self, exogenous-self, endogenous-other, and exogenous-other) versus the nonsocial control. Each bar shows mean decoding accuracy, error bars show SEM, red horizontal line shows chance level decoding. Significance indicated by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$) based on permutation testing (all but one of the significant p values also survived FDR correction for multiple comparisons across all six ROIs).

		L TPJ	R TPJ	L STS	R STS	MPFC	Precuneus
Endo-Self vs. nonsocial	Mean accuracy	58.50%	56.70%	53.00%	50.90%	55.20%	54.60%
	95% CI	55.5–61.6	53.5–59.8	49.3–56.8	47.5–54.0	52.3–58.0	51.5–57.7
	p value	0.0001*	0.0001*	0.0338*	0.2703	0.0026*	0.0027*
Exo-Self vs. nonsocial	Mean accuracy	60.70%	56.50%	56.30%	53.60%	54.50%	55.70%
	95% CI	57.0–64.4	53.7–58.8	53.5–59.0	50.5–56.6	51.7–57.1	51.9–59.4
	p value	0.0001*	0.0002*	0.0001*	0.0179*	0.0044*	0.0001*
Endo-Other vs. nonsocial	Mean accuracy	58.40%	52.70%	53.50%	52.20%	53.40%	55.40%
	95% CI	55.9–60.9	49.6–55.9	51.1–55.9	48.8–55.0	49.5–57.7	52.6–58.2

	p value	0.0001*	0.0497	0.0147*	0.0969	0.0219*	0.0014*
Exo- Other vs. nonsocial	Mean accuracy	56.70%	54.20%	53.10%	51.50%	53.40%	55.20%
	95% CI	53.8– 59.8	51.1– 57.3	50.4– 55.9	49.1– 54.0	49.8– 56.7	52.0–58.5
	p value	0.0002*	0.0065*	0.0319*	0.1901	0.0197*	0.0012*

Table 6. Decoding social versus nonsocial stories within the six ROIs. Mean decoding accuracy (%), 95% confidence interval (based on bootstrap distribution), and p value (based on permutation testing) are shown for each of the six ROIs. Results shown for each of four social story conditions (endogenous-self, exogenous-self, endogenous-other, and exogenous other) versus the nonsocial control. * indicates significant p values that survived correction for multiple comparisons across all six ROI (FDR-corrected $p < 0.05$).

Discussion

This study analyzed brain activity while people read stories about agents attending to objects in the environment. We examined whether specific brain areas could decode information about the type of attention referenced in the story (exogenous versus endogenous), and about the type of agent in the story (whether the agent was the subject reading the story or a different hypothetical person). We hypothesized that if the brain constructs a model of attentional states that is used in social cognition, then areas of the brain known to be involved in social cognition should be able to distinguish between the two types of attention, exogenous and endogenous, represented in the stories. Despite the extremely subtle difference between story conditions, our main analysis confirmed the hypothesis: the left TPJ showed significant decoding of information about endogenous versus exogenous attention.

These results support a new and growing body of evidence that the human brain constructs a model of attention to aid in theory of mind (Kelly et al., 2014; Pesquita et al., 2016; Vernet et al., 2019). The model includes information about attention that is deeper and more

complex than just gaze direction or an identification of the attended object. At least one aspect of attention incorporated into the model appears to be the manner in which attention moves to an object, whether it is internally directed, or externally induced. The processing of the model appears to engage the theory-of-mind cortical network. The left TPJ showed the strongest decoding result.

Why should the TPJ in particular have shown involvement in decoding someone else's attention state, rather than areas in the STS that are known for responding to the gaze direction of others (Marquardt et al., 2017)? As noted in the Introduction, we suggest that modeling someone else's attention, and processing someone else's gaze direction, are not the same. Gaze is only one of many cues that can be integrated to constrain a model of someone else's attention. In a study by Kelly and colleagues (2014) participants judged the attention state of a cartoon, and the TPJ was not found to be active in association with gaze direction, or in association with facial expression alone; however, both left and right TPJ were active during the integration of the two cues, gaze and expression, in order to judge the cartoon's attentional state. The current results involving significant decoding in the left TPJ is consistent with that prior finding.

It is not clear why the left hemisphere showed stronger activity than the right in the present task. Social cognition tasks often activate the TPJ bilaterally, but typically engage the right TPJ more (Saxe & Wexler, 2005). One speculation is that some aspect of the present task, perhaps explicitly instructing people that the task was a test of reading comprehension, caused an emphasis on linguistic processing, biasing the activity toward the left hemisphere. Other explanations for the left hemisphere bias may also be possible.

One alternative interpretation of the present finding in the TPJ is that people simulated attention orienting when reading the stories and activated the corresponding dorsal (endogenous)

and ventral (exogenous) attention networks in a ‘mirror-neuron-like’ fashion. Under this hypothesis, the TPJ, as part of the ventral attention network, was activated when reading the exogenous stories, and the dorsal attention network (consisting of the frontal eye fields, intraparietal sulcus, and middle temporal complex) should be active in association with reading the endogenous stories (Corbetta et al., 2008; Corbetta & Shulman, 2002; Fox et al., 2006). However, a control analysis showed that none of these dorsal attention network regions significantly decoded information concerning the attention type in the stories, which is incompatible with the attention-orienting simulation account of the TPJ result. Our findings are thus more consistent with the notion that the TPJ is involved in constructing a rich, perhaps implicit model of attention that may assist in social cognition which our lab has previously suggested (Graziano & Kastner, 2011; Kelly et al., 2014).

In addition to the endogenous-versus-exogenous comparisons, we also analyzed brain areas involved in self-versus-other encoding. We found evidence of self-versus-other encoding in the left TPJ, left STS, MPFC, and precuneus. The MPFC and precuneus have been previously implicated in self processing (Northoff et al., 2006; Ochsner et al., 2004; Passingham et al., 2010; Qin & Northoff, 2011; van Veluw & Chance, 2014), and the TPJ is consistently activated in fMRI studies involving self-recognition (van Veluw & Chance, 2014) and first-person perspective taking (Ionta et al., 2011). Although results from Chapter 3 may seem to contradict this finding, they are actually not at odds. Chapter 3 provides evidence that the underlying mechanisms for modeling attention for the self and others are shared, and these findings reiterate that the same areas are involved in each of these processes. Naturally, representations for the self and others would be distinct, but these results give further evidence that they are closely related. These results lend confidence to the present paradigm, showing that even the very subtle

differences between our story stimuli were able to reveal cortical results consistent with previous studies.

Contrary to prediction 3, we found no evidence for an interaction between attention type and agent type decoding in the theory-of-mind ROIs. In an exploratory searchlight analysis, we also found no evidence of an interaction effect in any other brain area. Although it is possible that our paradigm was simply not sensitive enough to detect subtle interaction effects, these results suggest that the brain encodes information about attention type in a similar manner in the self and in others. In light of previous results showing that the attribution of sensory awareness to others and to oneself have a shared representation in the TPJ (Kelly et al., 2014), we directly tested this notion in a post hoc cross-classification analysis focusing on the left TPJ, which was the only region showing significant decoding of attention type. We found that endogenous-versus-exogenous decoding significantly generalized across self-stories and other-stories, suggesting that there is an overlap in brain mechanisms that participate in the encoding of attention for others and in encoding of attention for the self.

Finally, significant decoding of the social-versus-nonsocial distinction was obtained across most of the theory-of-mind ROIs. This finding confirmed the validity of the paradigm and was expected based on previous experiments of the theory-of-mind network (Gallagher et al., 2000; Saxe & Kanwisher, 2003; van Veluw & Chance, 2014; Vogeley et al., 2001).

The use of a story-reading paradigm allowed us to systematically manipulate the kind of attention represented in the stimulus while keeping other experimental factors close to identical. The endogenous and exogenous story versions differed only with respect to a few key words specifying the type of attention, while the rest of the stories were semantically the same. The absence of any univariate effect within the ROIs or anywhere else in the brain, even at a liberal

uncorrected threshold, confirm that the stimuli were well matched. The finding of brain areas that significantly decoded the type of attention, despite the distinction between endogenous and exogenous attention being subtle and task-irrelevant, suggests that the human brain automatically, and possibly also implicitly (Pesquita et al., 2016), constructs a model of an agents' attention that specifies at least some dynamic aspects of how that attention is moving around the scene.

Chapter 7 - Conclusion

This dissertation has investigated one branch of the Attention Schema Theory, the proposal that individuals construct a mental model of the awareness of others and that this model has overlapping mechanisms with the mental model of one's own awareness. Using the concepts and methodological approaches from theory of mind and social attention, this body of work has explored the ways individuals understand and reconstruct the awareness and attention of other people and how these processes are connected to the construction of their own attention and awareness.

This work contributes to the understanding of these concepts in multiple ways. The study presented in Chapter 2 offers evidence that the ways in which individuals use and combine information to assess others' awareness is closely tied to their own social cognitive ability. Chapters 3 and 4 provide evidence that an individual's attentional biases seem to transfer when modeling the attention of others demonstrating a tight relationship between understanding others' attention and one's own. Finally, the experiments presented in Chapters 5 and 6 suggest that areas of the brain responsible for cognition about others represent the content of another agent's awareness as well as details about the agent's attention.

The study discussed in Chapter 2 explored how individuals construct the awareness of others by asking individuals to assess others' awareness given a limited number of visual cues. With a small amount of information available, subjects could employ strategies given any combination of the cues present. It was found that individuals' social cognitive ability predicts cue use, and that those best at social cognition tend to integrate information from different cues together to inform their judgments of others' awareness.

Chapter 3 explored a particular consequence that would arise if models of awareness for the self and others rely on shared mechanisms. If models of an agent's attention are interconnected with the construction of one's own attention, how would an individual's attentional bias relate to their judgment of others' awareness? The study described in Chapter 3 found that individuals transfer their own subtle visuospatial bias when assessing the visuospatial judgments of others. When taking the perspective of another agent, subjects were faster at responding when the other person's response lined up with their own bias. This result directly follows if these processes are related but are difficult to explain if these models are independent of one another.

Chapter 4 extended the work presented in Chapter 3, by testing patients with hemispatial neglect on their ability to perform theory of mind. Chapter 3 showed a relationship between individuals' visuospatial attention and their judgments of others' attention. If a subtle attentional bias could influence judgments of others, it seemed likely that a larger deficit (such as the kind an individual experiencing hemispatial neglect has) would show even more clearly the relationship between modeling one's own and another agent's attention and awareness. Neglect patients showed an overall deficit in social cognition.

Historically, theory of mind research using fMRI has equated differences in neural activity between true and false beliefs as evidence that people represent the content of others' minds. Chapter 5 explored this assumption by using a task that was sensitive enough to capture both a comparison between true and false beliefs as well as a comparison between two conditions with different content. The results suggested that right TPJ uniquely represents different content for others' minds, but the results showed no evidence to support the hypothesis that the right TPJ carried out social cognition by differently representing true and false beliefs.

Finally, chapter 6 further explored the rich models individuals make of others' minds. Chapter 6 investigated the extent to which individuals represent the details of others' attention. When modeling attention, do individuals merely track the object of attention or do they also represent details such as how attention is deployed? This experiment found that for both others and for themselves, individuals uniquely represent whether an agent engages their attention endogenously or exogenously.

Through behavioral studies in both neurotypical subjects and patients with hemispatial neglect, as well as fMRI experiments in neurotypical subjects, we found that the way people model their attention and awareness is closely tied to how they model another person's attention and awareness. This body of work provides evidence that the model individuals build for others' attention and awareness must be more complex and richer than merely a code for where someone else's gaze is pointing or whether their beliefs are true or false; and that the model is apparently linked to, and possibly even dependent on, one's own mechanisms for attention and awareness. This insight is important in determining a relationship between concepts that have heretofore been considered in isolation, but valuable work remains to be done mapping their underlying shared neural mechanisms. Future work should explore the neural underpinnings that are shared and determine dependencies they have in common. In sum, these studies provide support for the concept that the way people model the awareness of others as well as the way they model it for themselves have shared mechanisms and may also share an underlying neural substrate.

References

- Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8:14. DOI: <https://doi.org/10.3389/fninf.2014.00014>
- Adams, R. B., Jr, Albohn, D. N., & Kveraga, K. (2017). Social Vision: Applying a Social-Functional Approach to Face and Expression Perception. *Current Directions in Psychological Science*, 26(3), 243–248. <https://doi.org/10.1177/0963721417706392>
- Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporo-parietal Junction Activity in Theory-of-Mind Tasks: Falseness, Beliefs, or Attention. *Journal of Cognitive Neuroscience*, 21(6), 1179–1192. <https://doi.org/10.1162/jocn.2009.21082>
- Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences*, 4(7), 267–278. [https://doi.org/10.1016/S1364-6613\(00\)01501-1](https://doi.org/10.1016/S1364-6613(00)01501-1)
- Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7(4), 268–277. <https://doi.org/10.1038/nrn1884>
- Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with crosscorrelation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12:26–41. DOI:<https://doi.org/10.1016/j.media.2007.06.004>
- Baars, Bernard J. (1988). *A Cognitive Theory of Consciousness* (Cambridge, MA: Cambridge University Press).

- Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (2014). Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience*, 9(6), 817–824. <https://doi.org/10.1093/scan/nst048>
- Baron-Cohen, S. (1989). Perceptual role taking and protodeclarative pointing in autism. *British Journal of Developmental Psychology*, 7(2), 113–127.
- Baron-Cohen, S. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press. DOI: [https://doi.org/10.1016/S0166-2236\(96\)60013-7](https://doi.org/10.1016/S0166-2236(96)60013-7)
- Baron-Cohen, S. (2001). Theory of mind in normal development and autism. *Prisme*, 34(1), 74–183.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition*, 21(1), 37–46. [https://doi.org/10.1016/0010-0277\(85\)90022-8](https://doi.org/10.1016/0010-0277(85)90022-8)
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the Mind in the Eyes” Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, 42(2), 241–251. <https://doi.org/10.1111/1469-7610.00715>
- Baron-Cohen, S., Wheelwright, S., & Jolliffe, T. (1997). Is There a “Language of the Eyes”? Evidence from Normal Adults, and Adults with Autism or Asperger Syndrome. *Visual Cognition*, 4(3), 311–331. <https://doi.org/10.1080/713756761>
- Bartsch, K., & Estes, D. (1996). Individual differences in children’s developing theory of mind and implications for metacognition. *Learning and Individual Differences*, 8(4), 281–304.
- Bayliss, A. P., & Tipper, S. P. (2005). Gaze and arrow cueing of attention reveals individual differences along the autism spectrum as a function of target context. *British Journal of Psychology*, 96(1), 95–114.

- Bayliss A. P., Schuch S., & Tipper S. P. (2010). Gaze cueing elicited by emotional faces is influenced by affective context. *Visual Cognition*, 18(8):1214-1232.
DOI:10.1080/13506285.2010.484657
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289–300. DOI:<https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bio, B. J., Webb, T. W., & Graziano, M. S. (2018). Projecting one's own spatial bias onto others during a theory-of-mind task. *Proceedings of the National Academy of Sciences of the United States of America*, 115(7), E1684–E1689.
- Bio, B.J., Guterstam, A., Pinsk, M., Wilterson, A.I. & Graziano, M.S.A. (2021). Right Temporoparietal Junction Encodes More Information About Mental Experience of Others Than About False Versus True Beliefs. *BioRxiv*. DOI: <https://doi.org/10.1101/2021.06.10.447930>
- Birmingham, E., & Kingstone, A. (2009), Human Social Attention. *Annals of the New York Academy of Sciences*, 1156: 118-140. <https://doi.org/10.1111/j.1749-6632.2009.04468.x>
- Blanke, O., Mohr, C., Michel, C. M., Pascual-Leone, A., Brugger, P., Seeck, M., ... & Thut, G. (2005). Linking out-of-body experience and self processing to mental own-body imagery at the temporoparietal junction. *Journal of Neuroscience*, 25(3), 550-557.
- Block, N. (2009). Comparing the major theories of consciousness. In M. S. Gazzaniga, E. Bizzi, L. M. Chalupa, S. T. Grafton, T. F. Heatherton, C. Koch, J. E. LeDoux, S. J. Luck, G. R. Mangan, J. A. Movshon, H. Neville, E. A. Phelps, P. Rakic, D. L. Schacter, M. Sur, & B. A. Wandell (Eds.), *The cognitive neurosciences* (pp. 1111–1122). Massachusetts Institute of Technology.

- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10:433–436.
- Bristow, D., Rees, G., & Frith, C. D. (2007). Social interaction modifies neural response to gaze shifts. *Social Cognitive and Affective Neuroscience*, 2(1), 52-61.
- Brooks J. L., Della Sala S., & Darling S. (2014). Representational pseudoneglect: A review. *Neuropsychology Review*, 24:148–165.
- Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112(2), 337–342.
- Butterworth, G., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9(1), 55–72.
- Buxbaum, L. J., Ferraro, M. K., Veramonti, T., Farne, A., Whyte, J., Ladavas, E., Frassinetti, F., & Coslett, H. B. (2004). Hemispatial neglect: Subtypes, neuroanatomy, and disability. *Neurology*, 9;62(5):749-56. DOI:10.1212/01.wnl.0000113730.73031.f4.
- Calder, A. J., Beaver, J. D., Winston, J. S., Dolan, R. J., Jenkins, R., Eger, E., & Henson, R. N. A. (2007). Separate coding of different gaze directions in the superior temporal sulcus and inferior parietal lobule. *Current Biology : CB*, 17(1), 20–25. PubMed.
<https://doi.org/10.1016/j.cub.2006.10.052>
- Calder, A. J., Lawrence, A. D., Keane, J., Scott, S. K., Owen, A. M., Christoffels, I., & Young, A. W. (2002). Reading the mind from eye gaze. *Neuropsychologia*, 40(8):1129-38. doi: 10.1016/s0028-3932(02)00008-8
- Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, i–174.

- Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, 70(2):205-18. DOI:10.1037//0022-3514.70.2.205.
- 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(3), 314–325. <https://doi.org/10.1006/nimg.2000.0612>
- Charman, T. (2004). Matching Preschool Children with Autism Spectrum Disorders and Comparison Children for Language Ability: Methodological Challenges. *Journal of Autism and Developmental Disorders*, 34(1), 59–64. <https://doi.org/10.1023/B:JADD.0000018075.77941.60>
- Chiu, M. M., & Kuo, S. W. (2009). Social metacognition in groups: Benefits, difficulties, learning and teaching. In *Metacognition: New research developments* (pp. 117–136). Nova Science Publishers, Inc.
- Corbetta, M. (2014). Hemispatial neglect: Clinic, pathogenesis, and treatment. *Seminars in Neurology*, 34:514–523.
- Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, 58:306–324. DOI:<https://doi.org/10.1016/j.neuron.2008.04.017>
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3:201–215. DOI:<https://doi.org/10.1038/nrn755>
- Corcoran, R., Mercer, G., & Frith, C. D. (1995). Schizophrenia, symptomatology and social inference: Investigating “theory of mind” in people with schizophrenia. *Schizophrenia Research*, 17(1), 5–13.

- Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29:162–173.
DOI:<https://doi.org/10.1006/cbmr.1996.0014>
- D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., Maquet, P., & Salmon, E. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Cognitive Neuroscience*, 19(6), 935–944.
- David, N., Bewernick, B. H., Cohen, M. X., Newen, A., Lux, S., Fink, G. R., ... & Vogeley, K. (2006). Neural representations of self versus other: visual-spatial perspective taking and agency in a virtual ball-tossing game. *Journal of Cognitive Neuroscience*, 18(6), 898-910.
- Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist*, 13(6), 580–593.
- Descartes, R. (1985). *Principles of Philosophy*. In J. Cottingham, R. Stoothoff, & D. Murdoch (Trans.), *The Philosophical Writings of Descartes* (pp. 177-292). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511805042.007
- Döhnelt, K., Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., & Sommer, M. (2012). Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *NeuroImage*, 60(3), 1652–1661.
<https://doi.org/10.1016/j.neuroimage.2012.01.073>
- Driver IV, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, 6(5), 509–540.

- Ebert, S. (2015). Longitudinal Relations Between Theory of Mind and Metacognition and the Impact of Language. *Journal of Cognition and Development*, 16(4), 559–586.
<https://doi.org/10.1080/15248372.2014.926272>
- Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48(4), 384–392.
<https://doi.org/10.1037/0003-066X.48.4.384>
- Emery, N. J., Lorincz, E. N., Perrett, D. I., Oram, M. W., & Baker, C. I. (1997). Gaze following and joint attention in rhesus monkeys (*Macaca mulatta*). *Journal of Comparative Psychology*, 111(3), 286.
- Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16:111–116. DOI:<https://doi.org/10.1038/s41592-018-0235-4>
- Farrer, C., & Frith, C. D. (2002). Experiencing Oneself vs Another Person as Being the Cause of an Action: The Neural Correlates of the Experience of Agency. *NeuroImage*, 15(3), 596–603. <https://doi.org/10.1006/nimg.2001.1009>
- Farroni, T., Johnson, M. H., Brockbank, M., & Simion, F. (2000). Infants' use of gaze direction to cue attention: The importance of perceived motion. *Visual Cognition*, 7(6), 705–718.
<https://doi.org/10.1080/13506280050144399>
- Farroni, T., Massaccesi, S., Pividori, D., & Johnson, M. H. (2004). Gaze Following in Newborns. *Infancy*, 5(1), 39–60. https://doi.org/10.1207/s15327078in0501_2
- Fertuck, E. A., Jekal, A., Song, I., Wyman, B., Morris, M. C., Wilson, S. T., Brodsky, B. S., & Stanley, B. (2009). Enhanced 'Reading the Mind in the Eyes' in borderline personality

disorder compared to healthy controls. *Psychological Medicine*, 39(12):1979-88.

DOI:10.1017/S003329170900600X.

Fink, G. R., Marshall, J. C., Shah, N. J., Weiss, P. H., Halligan, P. W., Grosse-Ruyken, M., Ziemons, K., Zilles, K., & Freund, H.-J. (2000). Line bisection judgments implicate right parietal cortex and cerebellum as assessed by fMRI. *Neurology*, 54(6), 1324.

<https://doi.org/10.1212/WNL.54.6.1324>

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of “theory of mind” in story comprehension. *Cognition*, 57(2), 109–128. [https://doi.org/10.1016/0010-](https://doi.org/10.1016/0010-0277(95)00692-R)

[0277\(95\)00692-R](https://doi.org/10.1016/0010-0277(95)00692-R)

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102. DOI:[https://doi.org/10.1016/S1053-8119\(09\)](https://doi.org/10.1016/S1053-8119(09)47:S102)

Fox, M. D., Corbetta, M., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2006). Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proceedings of the National Academy of Sciences of the United States of America*, 103:10046–10051.

DOI:<https://doi.org/10.1073/pnas.0604187103>

Freeth, M., Foulsham, T., & Kingstone, A. (2013). What Affects Social Attention? Social Presence, Eye Contact and Autistic Traits. *PLoS ONE*, 8(1): e53286.

<https://doi.org/10.1371/journal.pone.0053286>

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, 5(3), 490–495.

<https://doi.org/10.3758/BF03208827>

- Friesen, C. K., & Kingstone, A. (2003). Abrupt onsets and gaze direction cues trigger independent reflexive attentional effects. *Cognition*, 87(1), B1–B10.
[https://doi.org/10.1016/S0010-0277\(02\)00181-6](https://doi.org/10.1016/S0010-0277(02)00181-6)
- Friesen, C. K., Ristic, J., & Kingstone, A. (2004). Attentional Effects of Counterpredictive Gaze and Arrow Cues. *Journal of Experimental Psychology: Human Perception and Performance*, 30(2), 319–329. <https://doi.org/10.1037/0096-1523.30.2.319>
- Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4), 694–724. PubMed. <https://doi.org/10.1037/0033-2909.133.4.694>
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2:189–210. DOI:<https://doi.org/10.1002/hbm.460020402>
- Frith, C. (2002). Attention to action and awareness of other minds. *Consciousness and Cognition*, 11(4), 481–487. [https://doi.org/10.1016/S1053-8100\(02\)00022-3](https://doi.org/10.1016/S1053-8100(02)00022-3)
- Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480), 671–678. <https://doi.org/10.1098/rstb.2006.2003>
- Frith, C. D., & Frith, U. (2006). The Neural Basis of Mentalizing. *Neuron*, 50(4), 531–534. <https://doi.org/10.1016/j.neuron.2006.05.001>
- Frith U., & Frith C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 358(1431):459-73. doi: 10.1098/rstb.2002.1218
- Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of “theory of mind”. *Trends in Cognitive Sciences*, 7(2), 77–83. [https://doi.org/10.1016/S1364-6613\(02\)00025-6](https://doi.org/10.1016/S1364-6613(02)00025-6)

- Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, 38(1), 11–21. [https://doi.org/10.1016/S0028-3932\(99\)00053-6](https://doi.org/10.1016/S0028-3932(99)00053-6)
- Gallup, A. C., Chong, A., & Couzin, I. D. (2012). The directional flow of visual information transfer between pedestrians. *Biology Letters*, 8(4), 520–522. <https://doi.org/10.1098/rsbl.2012.0160>
- Gallup, A. C., Hale, J. J., Sumpter, D. J. T., Garnier, S., Kacelnik, A., Krebs, J. R., & Couzin, I. D. (2012). Visual attention and the acquisition of information in human crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 109(19), 7245. <https://doi.org/10.1073/pnas.1116141109>
- German, T. P., Niehaus, J. L., Roarty, M. P., Giesbrecht, B., & Miller, M. B. (2004). Neural Correlates of Detecting Pretense: Automatic Engagement of the Intentional Stance under Covert Conditions. *Journal of Cognitive Neuroscience*, 16(10), 1805–1817. <https://doi.org/10.1162/0898929042947892>
- Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two Takes on the Social Brain: A Comparison of Theory of Mind Tasks. *Journal of Cognitive Neuroscience*, 19(11), 1803–1814. <https://doi.org/10.1162/jocn.2007.19.11.1803>
- Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *NeuroReport*, 6(13).
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data

processing framework in Python. *Frontiers in Neuroinformatics*, 5:13.

DOI:<https://doi.org/10.3389/fninf.2011.00013>

Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The Autism Diagnostic Observation

Schedule: Revised Algorithms for Improved Diagnostic Validity. *Journal of Autism and Developmental Disorders*, 37(4), 613. <https://doi.org/10.1007/s10803-006-0280-1>

Graziano, M. S. A. (2013). *Consciousness and the Social Brain*. Oxford University Press.

Graziano, M. S. A. (2019). Attributing awareness to others: the attention schema theory and its relationship to behavioral prediction. *Journal of Consciousness Studies: Controversies in Science & the Humanities*, 26:17–37.

Graziano, M. S. A., & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neuroscience*, 2(2), 98–113. PubMed.

<https://doi.org/10.1080/17588928.2011.565121>

Graziano, M. S. A., & Webb, T. W. (2014). A Mechanistic Theory of Consciousness.

International Journal of Machine Consciousness, 06(02), 163–176.

<https://doi.org/10.1142/S1793843014400174>

Graziano, M. S., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, 6, 500.

Greig, T. C., Bryson, G. J., & Bell, M. D. (2004). Theory of Mind Performance in

Schizophrenia: Diagnostic, Symptom, and Neuropsychological Correlates. *The Journal of Nervous and Mental Disease*, 192(1).

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48:63–72.

DOI:<https://doi.org/10.1016/j.neuroimage.2009.06.060>

- Grèzes, J., Frith, C. D., & Passingham, R. E. (2004). Inferring false beliefs from the actions of oneself and others: an fMRI study. *Neuroimage*, 21, 744-750.
- Guterstam, A., Bio, B. J., Wilterson, A. I., & Graziano, M. (2021). Temporo-parietal cortex involved in modeling one's own and others' attention. *ELife*, 10, e63551.
<https://doi.org/10.7554/eLife.63551>
- Halligan P. W., Fink G. R., Marshall J. C., & Vallar G. (2003). Spatial cognition: Evidence from visual neglect. *Trends in Cognitive Sciences*, 7:125–133
- Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P. D., & Maguire, E. A. (2009). Decoding neuronal ensembles in the human Hippocampus. *Current Biology*, 19:546–554.
DOI:<https://doi.org/10.1016/j.cub.2009.02.033>
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425–2430. DOI: <https://doi.org/10.1126/science.1063736>
- Hayward, D. A., & Ristic, J. (2013). Measuring attention using the Posner cuing paradigm: The role of across and within trial target probabilities. *Frontiers in Human Neuroscience*, 7.
<https://doi.org/10.3389/fnhum.2013.00205>
- Hebart, M. N., Gorgen, K., & Haynes, J. D. (2015). The decoding toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, 8:88. DOI: <https://doi.org/10.3389/fninf.2014.00088>
- Heber I. A., Siebertz S., Wolter M., Kuhlen T., & Fimm B. (2010). Horizontal and vertical pseudoneglect in peri- and extrapersonal space. *Brain and Cognition*, 73:160–166.
- Heilman, K. M., Bowers, D., Valenstein, E., & Watson, R. T. (1987). Hemispace and hemispatial neglect. In *Advances in Psychology*, (45)115-150.

- Heilman, K. M., & Valenstein, E. (1979). Mechanisms underlying hemispatial neglect. *Annals of Neurology*, 5:166–170.
- Hietanen, J. K., & Leppänen, J. M. (2003). Does Facial Expression Affect Attention Orienting by Gaze Direction Cues? *Journal of Experimental Psychology: Human Perception and Performance*, 29(6), 1228–1243. <https://doi.org/10.1037/0096-1523.29.6.1228>
- Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's Eyes Trigger Shifts of Visual Attention in Human Infants. *Psychological Science*, 9(2), 131–134. <https://doi.org/10.1111/1467-9280.00024>
- Hooker, C. I., Verosky, S. C., Germine, L. T., Knight, R. T., & D'Esposito, M. (2008). Mentalizing about emotion and its relationship to empathy. *Social Cognitive and Affective Neuroscience*, 3(3), 204–217. <https://doi.org/10.1093/scan/nsn019>
- Hori, E., Tazumi, T., Umeno, K., Kamachi, M., Kobayashi, T., Ono, T., & Nishijo, H. (2005). Effects of facial expression on shared attention mechanisms. *Physiology & Behavior*, 84(3), 397–405.
- Igelström, K. M., & Graziano, M. S. A. (2017). The inferior parietal lobule and temporoparietal junction: A network perspective. *Neuropsychologia*, 105:70–83.
- Igelström, K. M., Webb, T. W., & Graziano, M. S. (2015). Neural processes in the human temporoparietal cortex separated by localized independent component analysis. *Journal of Neuroscience*, 35(25), 9432–9445.
- Igelström, K. M., Webb, T. W., Kelly, Y. T., & Graziano, M. S. A. (2016). Topographical Organization of Attentional, Social, and Memory Processes in the Human Temporoparietal Cortex. *Eneuro*, 3(2), ENEURO.0060-16.2016. <https://doi.org/10.1523/ENEURO.0060-16.2016>

- Ionta, S., Heydrich, L., Lenggenhager, B., Mouthon, M., Fornari, E., Chapuis, D., Gassert, R., & Blanke, O. (2011). Multisensory mechanisms in temporo-parietal cortex support self-location and first-person perspective. *Neuron*, 70:363–374.
DOI:<https://doi.org/10.1016/j.neuron.2011.03.009>
- Jankowiak-Siuda, K., Baron-Cohen, S., Bialaszek, W., Dopierala, A., Kozłowska, A., & Rymarczyk, K. (2016). Psychometric evaluation of the “Reading the Mind in the Eyes” test with samples of different ages from a Polish population. *Studia Psychologica*, 58(1), 18.
- Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17:825–841. DOI: <https://doi.org/10.1006/nimg.2002.1132>
- Jewell G., & McCourt M. E. (2000). Pseudoneglect: A review and meta-analysis of performance factors in line bisection tasks. *Neuropsychologia*, 38:93–110.
- Kasari, C., Paparella, T., Freeman, S., & Jahromi, L. B. (2008). Language outcome in autism: Randomized comparison of joint attention and play interventions. *Journal of Consulting and Clinical Psychology*, 76(1), 125–137. <https://doi.org/10.1037/0022-006X.76.1.125>
- Kelly, Y. T., Webb, T. W., Meier, J. D., Arcaro, M. J., & Graziano, M. S. A. (2014). Attributing awareness to oneself and to others. *Proceedings of the National Academy of Sciences of the United States of America*, 111(13), 5012. <https://doi.org/10.1073/pnas.1401201111>
- Kingstone, A., Tipper, C., Ristic, J., & Ngan, E. (2004). The eyes have it!: An fMRI investigation. *Brain and Cognition*, 55(2), 269–271.
<https://doi.org/10.1016/j.bandc.2004.02.037>
- Kleiner M., Brainard D., & Pelli D. (2007). What’s new in Psychtoolbox-3? *Perception*, 36:1–16.

- Kobayashi, H., & Kohshima, S. (1997). Unique morphology of the human eye. *Nature*, 387:767–768. DOI:<https://doi.org/10.1038/42842>
- Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of the United States of America*, 110:5648–5653.
DOI:<https://doi.org/10.1073/pnas.1207992110>
- Koster-Hale, J., Bedny, M., & Saxe, R. (2014). Thinking about seeing: perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *Cognition*, 133:65–78. DOI: <https://doi.org/10.1016/j.cognition.2014.04.006>
- Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage*, 161:9–18. DOI:<https://doi.org/10.1016/j.neuroimage.2017.08.026>
- Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330:1830–1834.
- Kovács Á. M., 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(9): e106558. <https://doi.org/10.1371/journal.pone.0106558>
- Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103:3863–3868. DOI:<https://doi.org/10.1073/pnas.0600244103>
- Kristen, S., Sodian, B., Thoermer, C., & Perst, H. (2011). Infants' joint attention skills predict toddlers' emerging mental state language. *Developmental Psychology*, 47(5), 1207–1219.
<https://doi.org/10.1037/a0024808>

- Kuhn, D. (2000). Theory of mind, metacognition, and reasoning: A life-span perspective. In *Children's reasoning and the mind*. (pp. 301–326). Psychology Press/Taylor & Francis (UK).
- Kynast, J., Quinque, E. M., Polyakova, M., Luck, T., Riedel-Heller, S. G., Baron-Cohen, S., ... & Schroeter, M. L. (2020). Mindreading from the eyes declines with aging—evidence from 1,603 subjects. *Frontiers in Aging Neuroscience*, 12, 368.
- Laidlaw, K. E. W., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Potential social interactions are important to social attention. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14), 5548–5553. PubMed.
<https://doi.org/10.1073/pnas.1017022108>
- Langton, S. R. H., & Bruce, V. (1999). Reflexive visual orienting in response to the social attention of others. *Visual Cognition*, 6(5), 541–567.
<https://doi.org/10.1080/135062899394939>
- Langton, S. R. H., & Bruce, V. (2000). You must see the point: Automatic processing of cues to the direction of social attention. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2), 747–757. <https://doi.org/10.1037/0096-1523.26.2.747>
- Langton, S. R., Watt, R. J., & Bruce, I. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50-59. DOI:10.1016/s1364-6613(99)01436-9.
- Lee, J., Quintana, J., Nori, P., & Green, M. F. (2011). Theory of mind in schizophrenia: Exploring neural mechanisms of belief attribution. *Social Neuroscience*, 6(5–6), 569–581.
- Leekam, S. R., & Perner, J. (1991). Does the autistic child have a metarepresentational deficit? *Cognition*, 40(3), 203–218. [https://doi.org/10.1016/0010-0277\(91\)90025-Y](https://doi.org/10.1016/0010-0277(91)90025-Y)

- Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412:150–157. DOI:<https://doi.org/10.1038/35084005>
- Lord, C., & Jones, R. M. (2012). Annual Research Review: Re-thinking the classification of autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, 53(5), 490–509. <https://doi.org/10.1111/j.1469-7610.2012.02547.x>
- Marjoram, D., Job, D. E., Whalley, H. C., Gountouna, V. E., McIntosh, A. M., Simonotto, E., ... & Lawrie, S. (2006). A visual joke fMRI investigation into Theory of Mind and enhanced risk of schizophrenia. *Neuroimage*, 31(4), 1850-1858.
- Marquardt, K., Ramezanpour, H., Dicke, P. W., & Their, P. (2017). Following eye gaze activates a patch in the posterior temporal cortex that is not part of the human "Face Patch" System. *Eneuro*, 4:ENEURO.0317-16.2017. DOI: <https://doi.org/10.1523/ENEURO.0317-16.2017>
- Malhi, G. S., Lagopoulos, J., Das, P., Moss, K., Berk, M., & Coulston, C. M. (2008). A functional MRI study of Theory of Mind in euthymic bipolar disorder patients. *Bipolar Disorders*, 10(8), 943–956.
- Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, 62, 103–134.
- McDonald, S., Flanagan, S., Rollins, J., & Kinch, J. (2003). TASIT: A new clinical tool for assessing social perception after traumatic brain injury. *The Journal of Head Trauma Rehabilitation*, 18(3), 219–238. <https://doi.org/10.1097/00001199-200305000-00001>
- Meeren, H. K., van Heijnsbergen, C. C., & de Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45), 16518-16523.

- Milner A. D., Brechmann M., Pagliarini L. (1992). To halve and to halve not: An analysis of line bisection judgements in normal subjects. *Neuropsychologia*, 30:515–526.
- Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, 53:103–118. DOI:<https://doi.org/10.1016/j.neuroimage.2010.05.051>
- Mitchell, J. P. (2008). Activity in Right Temporo-Parietal Junction is Not Selective for Theory-of-Mind. *Cerebral Cortex*, 18(2), 262–271. <https://doi.org/10.1093/cercor/bhm051>
- Mitchell, J. P. (2009). Inferences about mental states. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 364, 1309-1316.
- Moore, C., & Dunham, P. J. (Eds.). (1995). *Joint attention: Its origins and role in development*. (pp. vii, 286). Lawrence Erlbaum Associates, Inc.
- Morales, M., Mundy, P., & Rojas, J. (1998). Following the direction of gaze and language development in 6-month-olds. *Infant Behavior & Development*, 21(2), 373–377. [https://doi.org/10.1016/S0163-6383\(98\)90014-5](https://doi.org/10.1016/S0163-6383(98)90014-5)
- Mundy, P., Block, J., Delgado, C., Pomares, Y., Vaughan Van Hecke, A., & Parlade, M. V. (2007). Individual Differences and the Development of Joint Attention in Infancy. *Child Development*, 78(3), 938–954. <https://doi.org/10.1111/j.1467-8624.2007.01042.x>
- Mundy, P., & Newell, L. (2007). Attention, Joint Attention, and Social Cognition. *Current Directions in Psychological Science*, 16(5), 269–274. <https://doi.org/10.1111/j.1467-8721.2007.00518.x>
- Mundy, P., & Sigman, M. (1989). The theoretical implications of joint-attention deficits in autism. *Development and Psychopathology*, 1(3), 173–183. <https://doi.org/10.1017/S0954579400000365>

- Mundy, P., Sigman, M., & Kasari, C. (1994). Joint attention, developmental level, and symptom presentation in autism. *Development and Psychopathology*, 6(3), 389–401.
<https://doi.org/10.1017/S0954579400006003>
- Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82:591–605.
DOI:<https://doi.org/10.1111/j.1469-185X.2007.00027.x>
- Nook, E. C., Lindquist, K. A., & Zaki, J. (2015). A new look at emotion perception: Concepts speed and shape facial emotion recognition. *Emotion*, 15(5), 569.
- Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—A meta-analysis of imaging studies on the self. *NeuroImage*, 31(1), 440–457. <https://doi.org/10.1016/j.neuroimage.2005.12.002>
- Nummenmaa, L., & Calder, A. J. (2009). Neural mechanisms of social attention. *Trends in Cognitive Sciences*, 13(3):135-43. DOI:10.1016/j.tics.2008.12.006.
- Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S. C. (2004). Reflecting upon Feelings: An fMRI Study of Neural Systems Supporting the Attribution of Emotion to Self and Other. *Journal of Cognitive Neuroscience*, 16(10), 1746–1772. <https://doi.org/10.1162/0898929042947829>
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science (New York, N.Y.)*, 308(5719), 255–258. PubMed.
<https://doi.org/10.1126/science.1107621>
- Pärnamets, P., & Olsson, A. (2020). Integration of social cues and individual experiences during instrumental avoidance learning. *PLoS Computational Biology*, 16(9), e1008163.

- Parton, A., Malhotra, P., & Husain, M. (2004). Hemispatial neglect. *Journal of Neurology, Neurosurgery and Psychiatry*, 75(1):13-21.
- Passingham, R. E., Bengtsson, S. L., & Lau, H. C. (2010). Medial frontal cortex: From self-generated action to reflection on one's own performance. *Trends in Cognitive Sciences*, 14(1), 16–21. <https://doi.org/10.1016/j.tics.2009.11.001>
- Pecchinenda, A., Pes, M., Ferlazzo, F., & Zoccolotti, P. (2008). The combined effect of gaze direction and facial expression on cueing spatial attention. *Emotion*, 8(5), 628–634. <https://doi.org/10.1037/a0013437>
- Pehrs, C., Zaki, J., Schlottermeier, L. H., Jacobs, A. M., Kuchinke, L., & Koelsch, S. (2017). The temporal pole top-down modulates the ventral visual stream during social cognition. *Cerebral Cortex*, 27(1), 777-792.
- Pellicano, E. (2010). Individual differences in executive function and central coherence predict developmental changes in theory of mind in autism. *Developmental Psychology*, 46(2), 530–544. <https://doi.org/10.1037/a0018287>
- Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: The perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of Cognitive Neuroscience*, 16(10), 1706–1716. <https://doi.org/10.1162/0898929042947900>
- Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development*, 60(3), 688–700.

- Pesquita, A., Chapman, C. S., & Enns, J. T. (2016). Humans are sensitive to attention control when predicting others' actions. *Proceedings of the National Academy of Sciences of the United States of America*, 113(31), 8669–8674. <https://doi.org/10.1073/pnas.1601872113>
- Petty, R. E., Briñol, P., Tormala, Z. L., & Wegener, D. T. (2007). *The role of metacognition in social judgment*. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 254–284). The Guilford Press.
- Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, 2:67–70. DOI:<https://doi.org/10.1093/scan/nsm006>
- Posner, M. I. (1980). Orienting of Attention. *Quarterly Journal of Experimental Psychology*, 32(1), 3–25. <https://doi.org/10.1080/00335558008248231>
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526. <https://doi.org/10.1017/S0140525X00076512>
- Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 159(2), 271–295. <https://doi.org/10.1007/s11229-007-9208-3>
- Qin, P., & Northoff, G. (2011). How is our self related to midline regions and the default-mode network? *NeuroImage*, 57(3), 1221–1233. <https://doi.org/10.1016/j.neuroimage.2011.05.028>
- Repacholi, B. M., Meltzoff, A. N., Rowe, H., & Toub, T. S. (2014). Infant, control thyself: Infants' integration of multiple social cues to regulate their imitative behavior. *Cognitive Development*, 32, 46–57. <https://doi.org/10.1016/j.cogdev.2014.04.004>

- Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, 9(1), 1027.
<https://doi.org/10.1038/s41467-018-03399-2>
- Rothmayr, C., Sodian, B., Hajak, G., Döhnell, K., Meinhardt, J., & Sommer, M. (2011). Common and distinct neural networks for false-belief reasoning and inhibitory control. *NeuroImage*, 56, 1705-1713.
- Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: A PET investigation of agency. *Nature Neuroscience*, 4:546–550.
- Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology. Human Perception and Performance*, 36(5), 1255–1266. <https://doi.org/10.1037/a0018729>
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind.” *NeuroImage*, 19(4), 1835–1842.
[https://doi.org/10.1016/s1053-8119\(03\)00230-1](https://doi.org/10.1016/s1053-8119(03)00230-1)
- Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, 16(2), 235–239. <https://doi.org/10.1016/j.conb.2006.03.001>
- Saxe, R, Moran, J. M., Scholz, J., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience*, 1(3), 229–234. <https://doi.org/10.1093/scan/nsl034>
- Saxe, R, & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*, 43(10), 1391–1399.
<https://doi.org/10.1016/j.neuropsychologia.2005.02.013>

- Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature*, 253(5489), 265–266. <https://doi.org/10.1038/253265a0>
- Schneider, B., Pao, Y., & Pea, R. D. (2013). Predicting students' Learning Outcomes Using Eye-Tracking Data. Learning Analytics and Knowledge Conference.
- Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage*, 101, 268-275.
- Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One*, 4:e4869.
- Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, 42, 9–34. <https://doi.org/10.1016/j.neubiorev.2014.01.009>
- Shulman, G. L., Pope, D. L., Astafiev, S. V., McAvoy, M. P., Snyder, A. Z., & Corbetta, M. (2010). Right hemisphere dominance during spatial selective attention and target detection occurs outside the dorsal frontoparietal network. *Journal of Neuroscience*, 30:3640–3651. DOI:<https://doi.org/10.1523/JNEUROSCI.4085-09.2010>
- Sodian, B., & Frith, U. (2008). Metacognition, theory of mind, and self-control: The relevance of high-level cognitive processes in development, neuroscience, and education. *Mind, Brain, and Education*, 2(3), 111–113. <https://doi.org/10.1111/j.1751-228X.2008.00040.x>
- Sommer, M., Döhnelt, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. *NeuroImage*, 35(3), 1378–1384. <https://doi.org/10.1016/j.neuroimage.2007.01.042>

- Sommer, M., Meinhardt, J., Eichenmüller, K., Sodian, B., Döhnelt, K., & Hajak, G. (2010). Modulation of the cortical false belief network during development. *Brain Research*, 1354, 123–131.
- Southgate, V., Senju, A., & Csibra, G. (2007). Action Anticipation Through Attribution of False Belief by 2-Year-Olds. *Psychological Science*, 18(7), 587–592.
<https://doi.org/10.1111/j.1467-9280.2007.01944.x>
- Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, 21(3), 489–510.
- Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18(7), 580–586.
- Swettenham, J. G. (1990). *The autistic child's theory of mind: A computer-based investigation*. University of York.
- Symons, L. A., Hains, S. M., & Muir, D. W. (1998). Look at me: Five-month-old infants' sensitivity to very small deviations in eye-gaze during social interactions. *Infant Behavior and Development*, 21(3), 531–536.
- Szczepanski S. M., & Kastner S. (2013). Shifting attentional priorities: Control of spatial attention through hemispheric competition. *Journal of Neuroscience*, 33:5411–5421.
- Tamir, D. I., & Mitchell, J. P. (2013). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology: General*, 142(1), 151.
- Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences of the United States of America*, 114(23), 5982–5987.

- Tipples, J. (2008). Orienting to counterpredictive gaze and arrow cues. *Perception & Psychophysics*, 70(1), 77–87.
- Turner R., & Felisberti F. M. (2017). Measuring mindreading: A review of behavioral approaches to testing cognitive and affective mental state attribution in neurologically typical adults. *Frontiers in Psychology*, 8:47.
- Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 Bias correction. *IEEE Transactions on Medical Imaging*, 29:1310–1320. DOI:<https://doi.org/10.1109/TMI.2010.2046908>
- Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly Journal of Experimental Psychology*, 59:11, 1931-1949, DOI: 10.1080/17470210500416342
- Vallar, G., & Perani, D. (1986). The anatomy of unilateral neglect after right-hemisphere stroke lesions. A clinical/CT-scan correlation study in man. *Neuropsychologia*, 24:609–622.
- Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, 30(3), 829–858.
- van Veluw, S. J., & Chance, S. A. (2014). Differentiating between self and others: An ALE meta-analysis of fMRI studies of self-recognition and theory of mind. *Brain Imaging and Behavior*, 8(1), 24–38.
- Vernet, M., Japee, S., Lokey, S., Ahmed, S., Zachariou, V., & Ungerleider, L. G. (2019). Endogenous visuospatial attention increases visual awareness independent of visual discrimination sensitivity. *Neuropsychologia*, 128:297–304. DOI:<https://doi.org/10.1016/j.neuropsychologia.2017.08.015>

- Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N. J., Fink, G. R., & Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *Neuroimage*, 14(1), 170–181.
- Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience*, 16(5), 817–827.
- Vuilleumier, P. (2002). Facial expression and selective attention. *Current Opinion in Psychiatry*, 15(3), 291–300. <https://doi.org/10.1097/00001504-200205000-00011>
- Webb, T. W., & Graziano, M. S. A. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, 6, Article 500.
- Webb, T. W., Igelström, K. M., Schurger, A., & Graziano, M. S. A. (2016). Cortical networks involved in visual awareness independent of visual attention. *Proceedings of the National Academy of Sciences of the United States of America*, 113:13923–13928.
- Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, 15, 728-755.
- Wellman H. M., Cross D., & Watson J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72:655–684.
- Williams, J. H., Waiter, G. D., Perra, O., Perrett, D. I., & Whiten, A. (2005). An fMRI study of joint attention experience. *Neuroimage*, 25(1), 133-140.
- Wilterson, A.I., Nastase, S., Bio, B.J., Guterstam, A. & Graziano, M.S.A. (2021). Involvement of the right temporoparietal junction in attention control, unexpected events, and predictive modeling. *Proceedings of the National Academy of Sciences of the United States of America*. DOI: <https://doi.org/10.1073/pnas.2026099118>

- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103–128.
- Xiao, Y., Geng, F., Riggins, T., Chen, G., & Redcay, E. (2019). Neural correlates of developing theory of mind competence in early childhood. *NeuroImage*, 184, 707–716.
<https://doi.org/10.1016/j.neuroimage.2018.09.079>
- Young-Browne, G., Rosenfeld, H., & Horowitz, F. (1977). Infant Discrimination of Facial Expressions. *Child Development*, 48(2), 555-562. doi:10.2307/1128653
- Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition*, 35(1), 41–68.
- Zaki, J. (2013). Cue integration: A common framework for social cognition and physical perception. *Perspectives on Psychological Science*, 8(3), 296-312.
- Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20:45–57. DOI:<https://doi.org/10.1109/42.906424>