Chapter 9

Functional neuroimaging of theory of mind

Jorie Koster-Hale and Rebecca Saxe

Introduction

In the decade since the last edition of *Understanding Other Minds*, the number of papers that use human neuroimaging tools to investigate the neural basis of theory of mind (ToM) has exploded from four (described in Frith & Frith's 2000 chapter) to, as of 2013, well over 400. Studying ToM with neuroimaging **works**. Unlike many aspects of higher-level cognition, which tend to produce small and highly variable patterns of responses across individuals and tasks, ToM tasks generally elicit activity in an astonishingly robust and reliable group of brain regions. In fact, convergence on this answer came almost immediately. By 2000, Frith and Frith concluded that "studies in which volunteers have to make inferences about the mental states of others activate a number of brain areas, most notable the medial [pre]frontal cortex [(mPFC)] and temporo-parietal junction [(TPJ)]." These regions remain the focus of most neuroimaging studies of ToM and social cognition, more than a decade later (see Adolphs, 2009, 2010; Carrington & Bailey, 2009; Frith & Frith, 2012; and Van Overwalle, 2008 for some recent reviews). To our minds, this consensus is one of the most remarkable scientific contributions of human neuroimaging, and the one least foreshadowed by a century of animal neuroscience.

Nevertheless, most of the fundamental questions about *how* our brains allow us to understand other minds remain unanswered; we have mainly discovered where to look next. We hope that this gap means the next decade of neuroimaging ToM will be even more exciting than the last one. In this chapter, we offer a perspective on the contribution that neuroimaging has made to the science of ToM in the last decade, and some thoughts on the contribution that it could make in the next one. The chapter has three sections: "Theory of mind and the brain" reviews the existing evidence for a basic association between thinking about people's thoughts and feelings and activity in this group of brain regions. "A strong hypothesis" discusses some objections, both theoretical and empirical, to a strong interpretation of this association, and our responses. "Where next?" highlights newer approaches to functional imaging data, which we expect will contribute to the future neuroscience of ToM, and their strengths and limitations.

Theory of mind and the brain

Theory of mind brain regions

Over the course of development, human children make a remarkable discovery: other people have minds, both similar to and distinct from their own. Other people see the world from a different angle, have different desires and preferences, and acquire different knowledge and beliefs. Children learn that other people's minds contain representations of the world which are often true and reasonable, but which may be strange, incomplete, or even entirely false. These discoveries (i.e. "building a Theory of Mind") help children to make sense of some otherwise mystifying behaviors: why mom would eat broccoli, even though there is chocolate cake available (e.g. Repacholi & Gopnik 1997), or why she is looking for the milk in the fridge, even though dad just put it on the table (e.g. Wimmer & Perner 1983).

As readers of this volume know well, developmental psychologists historically focused on one key transition in this developmental process—when and how children come to understand false beliefs. Assessing understanding of false beliefs has been taken to be a good measure of ToM capacity because it requires a child to understand both that someone can maintain a representation of the world, and that this representation may not match the true state of reality or the child's own beliefs. In a standard version of the false belief task, children might see that while their mother thinks the milk is in the fridge (having put it there 5 minutes ago), it is now actually on the table. The children are asked: "Where will she look for the milk?" or "Why is she looking in the fridge?". Five-year-old children, like adults, usually predict that she will look in the fridge, because that is where she thinks the milk is (Wellman, Cross, & Watson, 2001). Three-year-olds, however, predict that she will look on the table, explaining that she wants the milk and the milk is on the table (at least when asked explicitly; see, e.g. Onishi & Baillargeon (2005), Saxe (in press), and Southgate, Senju, & Csibra (2007), for further discussion of ToM behavior in pre-verbal children). In fact, when three-year-olds see her look in the fridge instead, some will go so far as to fabricate belief-independent explanations, stating that she no longer wants the milk, and must be looking for something else (Wellman et al., 2001; Wimmer & Perner 1983).

Building off decades of experience in developmental psychology, the first neuroimaging studies of ToM also used versions of false belief tasks. Adults, lying in positron emission tomography (PET) or magnetic resonance imaging (MRI) scanners, read short stories describing a person's action (see Figure 9.2), and were asked to explain that action (usually silently to themselves, to avoid motion artifacts). These early studies revealed increased levels of blood oxygen and glucose uptake (indirect measures of metabolic activity, henceforth called "activity"), in a small, but consistent group of brain regions—left and right TPJ and mPFC, as noted by the Friths, and also medial parietal cortex (precuneus, PC) and more anterior regions of the superior temporal sulcus (STS), down to the temporal poles (Figure 9.1).

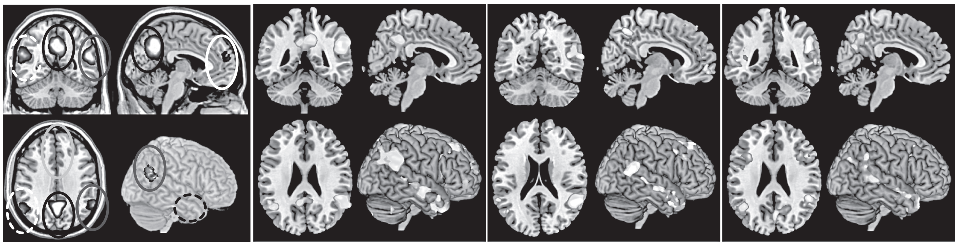


Figure 9.1 Brain regions commonly recruited in Theory of Mind tasks. (Left) Average activity in 63 subjects reading stories about false beliefs, compared to stories about false photographs ($P < 0.05$, corrected; see Dodell-Feder et al., 2011; Saxe & Kanwisher, 2003) overlaid on an average brain. Colored ellipses indicate standard locations of the right TPJ (red), left TPJ (yellow), right anterior STS (orange) medial parietal/precuneus (blue), and mPFC (green). Other three panels: activity in the same task in three example individual participants, overlaid on the same average brain anatomy for easy comparison ($P < 0.001$, uncorrected). Thanks to Nicholas Dufour for the images.

AQ1: Please change color indication in caption

Activity during the false belief task is, of course, far from sufficient evidence that these brain regions have any role in understanding other minds. We believe this proposition becomes more compelling after reviewing the range of different experimental tasks and paradigms that have been used successfully over the years, across laboratories and countries, each aiming to elicit some aspect of "understanding other minds." While some studies used complex verbal narratives, other researchers have used simple sentences or non-verbal cartoons; some studies explicitly instruct participants to think about a person's thoughts, and others have elicited ToM spontaneously. The heterogeneity of methods, materials, and participant demographics makes it especially impressive that these studies have converged on the same regions of the brain. In this section, we will review a sample of these different procedures.

In the original theory of mind functional MRI experiments, both the content of the materials and the explicit instructions focused participants' attention on (or away from) thinking about someone's mind. For example, in an early PET study, Fletcher Happe, Frith, Baker, Dolan, Frackowiak, et al. (1995) told participants that they would be reading different kinds of verbal passages. Just before each item, the participant was told what kind of story was coming next. If it was a "mental" story, participants were instructed that it was "vital to consider the thoughts and feelings of the characters," and then shown a story revolving around someone's mental states (see example in Figure 9.2). If the story was a "physical" story, participants were first instructed that thinking about thoughts and feelings was irrelevant and undesirable, then shown a control story (see example in Figure 9.2). After each story, participants were instructed to silently answer an action-explanation question, such as "Why did the prisoner say that?". Glucose consumption increased in the theory of mind brain regions while people read the mental stories, relative to the control stories.

The same design has been used with non-verbal stimuli. In an early fMRI experiment (Gallagher, Happé, Brunswick, Fletcher, Frith, & Frith 2000), participants both read the stories used by Fletcher et al. (1995) and were shown cartoons depicting visual jokes that either relied on ToM (in which understanding the joke depended on attribution of either a false belief or ignorance), or other types of humor (such as puns, idioms, and physical humor). Again, participants were cued in advance about whether to expect a mental or control cartoon (for examples, see Figure 9.2). For both types of cartoons, they were asked to silently contemplate the meaning; for mental cartoons, they were also explicitly instructed to consider the thoughts and feelings of the characters. With less than 20 minutes of scanning for each task, Gallagher and colleagues found that the same group of brain regions showed increased activity for both verbal and non-verbal ToM stimuli; these regions include the bilateral temporal-parietal junction and the middle prefrontal cortex. Similar convergence of the activity elicited by verbal and non-verbal stimuli has been found by Kobayashi, Glover & Temple (2007).

Sommer, Döhnel, Sodian, Meinhardt, Thoermer, & Hajak (2007) also non-verbal stimuli to focus participants' attention on the thoughts of a character, but without explicitly cuing the condition. They showed participants a series of cartoon images depicting a story—Betty hides her ball, Nick moves it, and then Betty comes back to look for her ball. In half of the trials, Betty looks into the box where she thinks the ball is (the expected condition); in the other half, she looks into the other box (the unexpected condition). Participants judged whether, based on the character's beliefs, the character's action was expected or unexpected. Rather than explicitly labeling the mental and control conditions, a key contrast was between the beginning of the trial, before mental state inferences were possible, and the end of the trial, when participants had presumably made belief inferences to complete the task. The second key contrast was between trials in which Betty had a false belief (so that predicting her action required considering her thoughts) and trials in

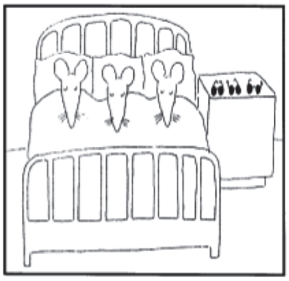
Mental	Control
During the war, the Red army capture a member of the Blue army. They want him to tell them where his armies' tanks are. They know that they are either by the sea or in the mountains. They know that the prisoner will not want to tell them, he will want to save his army, and so he will certainly lie to them. The prisoner is very brave and very clever, he will not let them find his tanks. The tanks are really in the mountains. Now when the other side ask him where his tanks are he says, "They are in the mountains."	Two enemy powers have been at war for a very long time. Each army has won several battles, but no the outcome could go either way. The forces are equally matched. However, the Blue army is stronger than the Yellow army in air foot soldiers and artillery. But the yellow army is stronger than the Blue army in air power. On the day of the final battle, which will decide the outcome of the war, there is heavy fog over the mountains where the fighting is about to occur. Low-lying clouds hang over the soldiers. By the end of the day, the Blue army has won.
	
Brad had no money, but just had to have the beautiful ruby ring for his wife. Seeing no salespeople around, he quietly made his way closer to the counter. He was seen running out the door.	While playing in the waves, Sarah's Frisbee went flying toward the rocks in the shallow water. While searching for it, she stepped on a piece of glass. Sarah had to wear a bandage on her foot for a week.
The path to the castle leads via the lake. But children tell the tourists: ''The way to the castle goes through the woods.'' The tourists now think that the castle is via the woods or lake?	The sign to the monastery points to the path through the woods. While playing the children make the sign point to the golf course. According to the sign the monastery is now in the direction of the golf course or woods?
How likely is Queen Elizabeth to think that keeping a diary is important?	How likely is Queen Elizabeth to sneeze when a cat is nearby?
John was on a hike with his girlfriend. He had an engagement ring in his pocket and at a beautiful overlook he proposed marriage. His girlfriend said that she could not marry him and began crying. John sat on a rock and looked at the ring.	Joe was playing soccer with his friends. He slid in to steal the ball away, but his cleat stuck in the grass and he rolled over his ankle, breaking his ankle and tearing the ligaments. His face was flushed as he rolled over.
That morning, people sat around looking at each other, wondering if they were dreaming, because everything looked purple. Some people were shocked. Some people thought that it was funny to see everybody all purple. But even the smartest scientists didn't know what had happened.	The whole world had turned purple overnight. Just about everything was purple, included the sky and the ocean and the mountains and the trees. The tallest skyscrapers and the tiniest ants were all purple. The bicycles and furniture and food were purple. Even the candy was purple.
In spite of her neighbourhood, Erica has a strong dislike of violence, and believes that conflicts can usually be resolved without fists.	Erica lives in Los Angeles. One night recently she was in a bar where a fight broke out between two drunk men and she was caught in between.
Sam thinks he can grow trees with fruit that taste like pizza. How likely is it that Sam wants these trees for a treehouse too?	In the backyard are trees with fruit that taste like pizza when ripe. How likely is it that these trees can be used for building a treehouse?

Figure 9.2 Samples of experiments that elicit thinking about thoughts and feelings by manipulating the content of the stimuli. Sample stimuli from Fletcher et al. (1995), Gallagher et al. (2000), Mason & Just (2010), Perner et al. (2006), Lombardo et al. (2010), Bruneau et al. (2012), Saxe et al. (2009), Saxe & Wexler (2005), Young et al. (2010b).

which Betty knew where the ball was all along (so her action could be predicted based on the actual location of the ball).[1] Both contrasts revealed activity in ToM brain regions.

Another way to endow non-verbal stimuli with mentalistic content is by altering the movements of simple geometric shapes (Heider & Simmel, 1944). For example, in a PET study, Castelli, Happé, Frith, & Frith (2000) showed participants animations of two triangles moving around. Participants were instructed that while some of the triangles "just move about with random movement [...] disconnected from each other" (the control condition), other animations would show "two triangles doing something more complex together, as if they are taking into account their reciprocal feelings and thoughts [...] for example, courting each other," (the mental condition). Participants watched the animations, and then described what the triangles were doing. ToM animations elicited more activity than the random animations in the TPJ, the nearby superior temporal sulcus (STS), and the mPFC.

Some experiments elicit thinking about thoughts simply by describing those thoughts in words. For example, participants can answer questions about people's mental characteristics, such as "How likely is Queen Elizabeth to think that keeping a diary is important?" vs. their physical traits, such as "How likely is Queen Elizabeth to sneeze when a cat is nearby?" (Lombardo, Chakrabarti, Bullmore, Wheelwright, Sadek, Suckling, et al., 2010; Mitchell, Macrae, & Banaji, 2006); or they can read single sentences describing thoughts ("He thinks that the nuts are rancid") or facts ("It is likely that the nuts are rancid"; Zaitchik, Walker, Miller, LaViolette, Feczko, & Dickerson, 2010). In both cases, the items related to mental states elicited more activity in ToM regions than the control conditions.

Other experiments elicit thinking about thoughts indirectly. Saxe & Kanwisher (2003) used verbal stories based on either inferences about false beliefs or about physical events, similar to Fletcher et al. (1995). However, participants were not given any explicit instructions about the different kinds of stories. In Experiment 1, participants did not give any response, while in Experiment 2, they responded to fill-in-the-blank questions about details in the stories. Similarly, Mason & Just (2010) had participants read short stories about actions, and then answer simple comprehension questions. Critically, the stories elicited spontaneous inferences about unstated, but implied, events; some of these inferences were about a character's thoughts (mental), and some about purely physical events (control). Compared with the original Fletcher et al. (1995) stories, the stories in these experiments (see examples in Figure 9.2) were shorter, and included less (or no) explicit description of thoughts and feelings (see also Bruneau, Pluta, & Saxe, 2011). Instead, the thoughts and feelings of the characters had to be inferred. Listening to the mental stories elicited strong activation in ToM regions relative to control stories, suggesting consideration of the character's thoughts despite the absence of explicit instruction.

Another procedure for eliciting spontaneous ToM in the scanner was developed by Spiers & Maguire (2006). Participants engaged in naturalistic actions (e.g. driving a taxicab through bustling London streets) in a rich virtual reality environment. After the scan and without prior warning, participants reviewed their performance, and were asked to recall their spontaneous thoughts

[1] In the original study, this contrast could have been due to a difference between false vs. true beliefs, or between representing a belief (required for the false trials) and making a prediction based solely the actual location of the ball (possible for the true beliefs). Subsequent work has shown that activation observed by Sommer et al. (2007) is due to the latter: individuals often simply reduce the information they need to process by choosing not to represent true beliefs as a mental state (Apperly et al., 2007), and when this is controlled for, neuroimaging has revealed indistinguishably high activation for true and false beliefs (Döhnel, Schuwerk, Meinhardt, Sodian, Hajak, & Sommer, 2012; Jenkins & Mitchell, 2010; Young et al., 2010b).

during each of the events. These recollections were coded for content concerning the thoughts and intentions of the taxicab customers and the other drivers and pedestrians on the road (e.g. "I reckon that she's going to change her mind") and used these coded events to predict neural responses. They found that when participants were thinking about someone else's intentions, but not during other events, regions in the ToM network showed increased activity.

Other studies have targeted spontaneous consideration of others' intentions by asking participants to make moral judgments. Morally relevant facts appear to rely in part on consideration of intentions (Cushman, 2008), and evoke increased activity in the ToM network, relative to other facts in a story (Young & Saxe, 2009a). The same brain regions are recruited when participants are forced to choose between acting on a personal desire vs. a conflicting moral principle, compared to deciding between two conflicting personal desires (Sommer, Rothmayr, Döhnel, Meinhardt, Schwerdtner, Sodian, et al., 2010). These regions are also recruited in children watching an animation of one person intentionally harming another, compared to animations of other painful and non-painful situations (Decety, Michalska, & Akitsuki, 2008).

In fact, when participants read a story, they appear to automatically represent the thoughts and feelings of the characters in order to make sense of the plot, even if instructed to perform an orthogonal task. For example, Koster-Hale & Saxe (2011) had participants read short verbal stories, and then make a delayed-match-to-sample judgment, indicating whether a single probe word occurred in the story (match) or not (non-match); half of the stories described a false belief and half described physical representations. Despite the word-level task (and no mention of ToM in the explicit instructions), the contrast revealed activation across the ToM network. Similarly, in two fMRI experiments with children aged 5–12 years, children heard child-friendly verbal stories, describing characters' thoughts and feelings vs. physical events; children answered orthogonal (delayed-match-to-sample) questions about each story. As with adults, we found increased activation in the ToM network when children were listening to stories involving thoughts and feelings (Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009; Gweon, Dodell-Feder, Bedny, & Saxe, 2012).

Thinking about thoughts and feelings can be also manipulated by changing the task while holding the stimuli constant (Figure 9.3). In an early PET study, Goel, Grafman, Sadato, & Hallett (1995) showed participants sets of 75 photographs of objects, some modern and familiar, and some from pre-fifteenth century North American aboriginal culture. Participants either judged whether the object was elongated along the principle axis (the control task) or whether "someone with the background knowledge of Christopher Columbus could infer the [object's] function" (the mental task). They found increased ToM activation when participants were considering Christopher Columbus, but not when making the physical judgments. Similarly, Walter and colleagues showed participants sequences of three cartoon images (Schnell, Bluschke, Konradt, & Walter (2011; Walter, Schnell, Erk, Arnold, Kirsch, Esslinger, et al., 2010). Participants either judged, on each picture, whether "the protagonist feels worse/equal/better, compared to the previous picture" (the mental task) or whether "the number of living beings [in the image is] smaller/equal/greater, compared to the previous picture" (the control task). In another set of studies, Baron-Cohen and others (Adams, Rule, Franklin Jr, Wang, Stevenson, Yoshikawa, 2010; Baron-Cohen & O'Riordan, 1999; Cohen, Wheelwright, & Hill, 2001; Platek, Keenan, Gallup, & Mohamed, 2004) showed participants pictures of a person's eyes. Participants pressed a button to indicate either the mental/emotional state of the person in the picture (e.g. *embarrassed*, *flirtatious*, *worried*; the mental task) or their gender (the control task). Mitchell, Banaji, & MacRae (2005) asked participants to either judge either how happy a person was to be photographed (mental) or how symmetric their face was (control). In all of these cases, the mental tasks activated the ToM regions more than the control tasks.

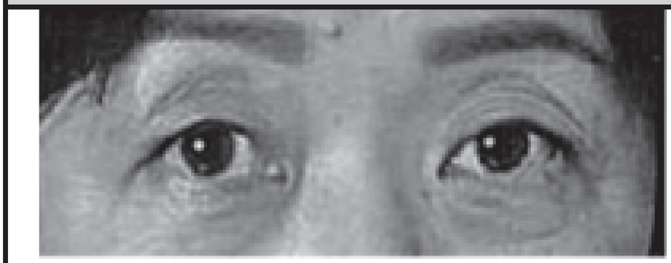
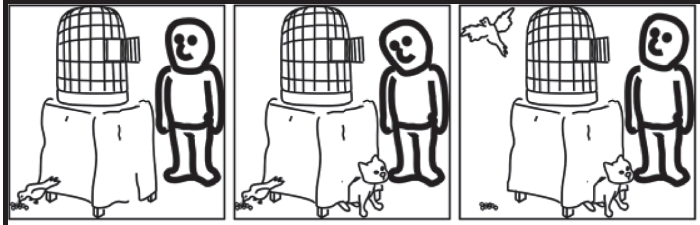
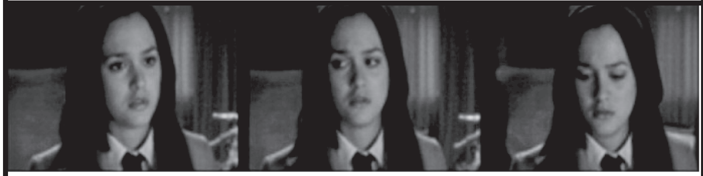
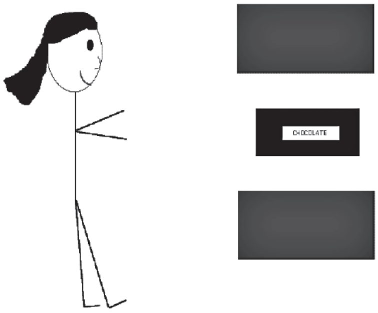
Stimulus	Mental Task
	Worried or friendly?
	(In each panel) Does he feel better or worse than in the previous panel?
	Why does she feel this way?
	Where does the girl think the chocolate is?

Figure 9.3 Samples of experiments that elicit thinking about thoughts and feelings by holding the stimulus constant, and manipulating the participants' task. Examples from Adams et al. (2010), Schnell et al. (2010), Spunt & Lieberman (2012), and Saxe, Schulz & Jiang (2006).

Spunt and colleagues have recently developed a clever paradigm for eliciting ToM using a simple task manipulation. In their first experiment, Spunt, Satpute, & Lieberman (2011) showed participants pictures of simple human actions (e.g. a person riding a bike), and instructed them to silently answer one of three questions: **why** the person is doing the action (e.g. to get exercise), **what** the person is doing (e.g. riding a bike), or **how** the person is doing it (e.g. holding handlebars). These questions require successively less consideration of the mind of the person and, correspondingly, showed successively less ToM region activity. Spunt & Lieberman (2012) replicated the result using a similar paradigm with brief naturalistic film clips of facial expressions of emotions. Participants judged either **how** the person is expressing their emotion (e.g. "looking down and away," the control task) or **why** she is feeling that emotion (e.g. "she is confused because a friend let her down," the mental task). Again, ToM regions were recruited more when thinking about why than how.

Using these types of parametric designs and analyzing the continuous magnitude of response in ToM regions may provide a powerful tool for studying the neural basis of ToM, especially in combination with computational models of ToM, which offer quantitative predictions of both when and how much (or how likely) people are thinking about others' thoughts. For example, Bhatt, Lohrenz, Camerer, & Montague (2010) created a competitive buying and selling game in which participants could try to bluff about the value of an object. The authors predicted that reliance on ToM would increase in proportion to the *riskiness* of the bluff, i.e. the discrepancy between the object's true value and the proposed price. Consistent with this idea, a region near the right TPJ showed activity correlated with bluff riskiness across trials. They suggested that participants may be more likely to engage in ToM, engage in more ToM, or engage in ToM for longer, when they are making a riskier bluff relative to a less risky bluff, and this relationship is continuous over a large range of possible risks.

Along with manipulating thinking about thoughts across stimuli and tasks, it is also possible to look at *when* participants are thinking about thoughts within a single ongoing stimulus and task. Stories about human actions and beliefs can be broken down into sections, separating the description of the background and set-up from the specific sentence that describes or suggests a character's mental states. Thinking about other minds can thus be pinned to a specific segment within an ongoing story. The right temporo-parietal junction, in particular, shows activity at the point within a single story when a character's thoughts are mentioned (Mason & Just, 2010; Saxe & Wexler, 2005; Young & Saxe, 2008). A similar manipulation, dividing a 60-second story into 20-second segments, only one of which has mental information, has been used in children (Saxe et al., 2009).

In sum, neuroimaging experiments on understanding other minds produce similar results, across a wide range of participants, methods, and materials. Similar experiments have been conducted in Britain, the USA, Japan, Germany, China, the Netherlands, and Italy (e.g. Anna Leshinskaya, personal communication; Moriguchi, Ohnishi, Lane, Maeda, Mori, Nemoto, et al., 2006; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Schnell et al., 2010; Markus van Ackeren, personal communication). The same brain regions are found in participants ranging from 5 years old (Decety et al., 2008; Gweon et al., 2012; Saxe et al., 2009) to at least 65 years old (e.g. Bedny, Pascual-Leone, & Saxe, 2009; Fletcher et al., 1995. These regions recruited in adults with high functioning autism (Dufour, Redcay, Young, Mavros, Moran, Triantafyllou et al., 2012), and adults who have been completely blind since birth (Bedny et al., 2009); ongoing work in our laboratory suggests they are found in congenitally deaf adults as well. As described above, the same set of regions responds whether the stimuli is presented in text or with pictures, visually or aurally. Participants can be thinking about the thoughts and feelings of a real person or a fictional character.[2] The stimuli can include complex narratives, or just a single thought; participants can be instructed to consider others' thoughts and feelings, or be led to do so spontaneously.

The range of tasks, stimuli, and populations make it all the more striking that these experiments converge on the same conclusions. A consistent group of brain regions shows increased metabolic activity across all of these experiments in the "mental" or "theory of mind" condition: regions in bilateral temporo-parietal junction, medial parietal/precuneus, medial prefrontal cortex, and anterior superior temporal sulcus. Though this generalization is striking on its own, a key question is: why? Which cognitive process, invoked by all of these diverse tasks, is specifically necessary and sufficient to elicit activity in these brain regions?

Specificity

During the initial discovery of the ToM brain regions, the first experiments (e.g. Fletcher et al., 1995; Gallagher et al., 2000) compared two conditions that differed on multiple dimensions. Compared with the control stories, the stories about false beliefs also included more individual characters, more specific human actions, more implied human emotions, more invisible causal mechanisms, more social roles, more unexpected events, more demand to consider false representations of the world, different syntax, and so on. In fact, an inherent risk of such complex stimuli is that there may be hidden dimensions along which the stimuli grouped into separate conditions differ and that it is these differences, rather than the intended manipulation, that lead to differential brain activity.

Given all these dimensions, how can we infer which are the necessary and sufficient features that led to activity in each region during a given task? One approach is to try to design an experiment that contrasts **minimal pairs**: stimuli and tasks that differ only in one key dimension, but are exactly identical on all other dimensions. Taking with approach, Saxe, Schulz, and Jiang (2006b) were able to match both the stimulus and the participant's response, using task instructions to change just how the participants *construed* the stimulus. The stimulus was a stick-figure animation of a girl. Modeled after a false-belief transfer (change of location) task, a bar of chocolate moved from one box to another, while the girl either faced toward the transfer or away. In the first half of the experiment, rather than introducing the task as a false-belief task, participants were trained to treat the stick-figure as a physical cue to the final location of the chocolate bar using three rules, including the critical Rule 1: 'Facing = last; Away = first. If the girl is facing the boxes at the end of the trial, press the button for the last box. If the girl is looking away from the boxes, press the button for the first box." Participants were accurate in the task, but found it difficult and unnatural. In the second half of the experiment, participants were then told that for Rule 1, another strategy was possible: namely to view the stick-figure as a person and to consider that person's thoughts. Rule 1 was equivalent to judging, based on what the character had seen, where she *thought* the chocolate was. Both ways of solving "Rule 1" generate the same behavioral responses, but only in the second half of the experiment were participants construing Rule 1 as referring to a character's thoughts. We found that, though the stimuli and the responses were identical across tasks, right TPJ activity was significantly higher in the second half, when participants were using ToM to solve the puzzle, rather than the simple association rule.

[2] Interestingly, participants can also be assigning hypothetical thoughts and preferences to themselves (e.g. Lombardo et al., 2010; Vogeley et al., 2001). Note, however, that not all metacognition elicits ToM activity. The link between attributing hypothetical thoughts to the self, vs. other kinds of metacognition, is not completely clear (see Saxe & Offen, 2009).

Another approach to dealing with the many dimensions of ToM stimuli is to systematically vary or match each of these dimensions in a long series of experiments. Although each experiment has many differences between the mental and control conditions, across the whole set of experiments, most other kinds of differences are eliminated, leaving only one systematic factor: thinking about thoughts.

For example, Gallagher et al. (2000) showed that none of the low-level features of the original verbal stimuli (e.g. number of nouns, number of straight edges, retinal position) are **necessary** to elicit activity in these brain regions, because they found the same patterns of activity in response to verbal false belief stories and non-verbal false belief cartoons. Within verbal stories, it is not necessary to explicitly state a character's thoughts or beliefs: there is activity in these regions both when people read about a character's thoughts and when they infer those thoughts from the character's actions (Mason & Just, 2010; Young & Saxe, 2009a). Nor is it necessary that the beliefs in question be false: true beliefs, false beliefs, and beliefs whose veracity is unknown are all sufficient to elicit robust activity in these brain regions (Döhnel, Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., & Sommer, 2012; Jenkins et al., 2010; Young, Nichols, & Saxe, 2010c).

Other experiments showed that the presence of a human character in the stimuli is not **sufficient**: stories that describe a character's physical appearance, or even their internal (but not mental) experiences, like hunger or queasiness or physical pain, elicit much less response than stories about the character's beliefs, desires, and emotions (Bedny et al., 2009; Bruneau et al., 2011; Lombardo et al., 2010; Saxe & Powell 2006). It is also not sufficient for a story to describe invisible causal mechanisms (like melting and rusting, Saxe & Kanwisher, 2003), unexpected events (like a ball of dough that rises to be as big as a house, Young, Dodell-Ferer & Saxe, 2010b; Gweon et al., 2012), or people's stable social roles (including kinship and professional relationships, Saxe & Wexler, 2005; Gweon et al., 2012), if the story does not also invoke thinking about a person's thoughts.

One particularly important dimension to test was whether considering any representation of the world, mental or otherwise, would be sufficient to elicit activity in these brain regions. Understanding other minds often requires the ability to suspend one's own beliefs and knowledge, and consider the world as it would seem from another perspective. These cognitive processes have been called **meta-representation** (the ability to conceive of distinct representations of the world, Aichhorn, Perner, Weiss, Kronbichler, Staffen, & Ladurner, 2009; Perner, 1991; Perner et al., 2006), and **decoupling** (the ability to suspend one's own knowledge in order to respond from a different perspective, Leslie & Frith, 1990; Liu, Sabbagh, & Gehring, 2004). Since meta-representation and decoupling are such essential ingredients of understanding other minds, and especially understanding false beliefs, many scientists initially hypothesized that brain regions recruited by false belief tasks most likely performed one of these two functions. To test this hypothesis, we need stimuli or tasks that require meta-representation and decoupling, but are not about understanding other minds. Currently, the best such example are stories about "false signs" and "false maps" (Zaitchik, 1990). Like beliefs, signs and maps represent (and sometimes misrepresent) reality. Thinking about the world as depicted in a map requires the capacity for meta-representation; when the map is wrong, reasoning about the world as it seems in the map requires decoupling from one's own knowledge of reality. Nevertheless, stories about false signs and maps elicit much less activity in these brain regions (especially right TPJ) than stories about beliefs (Aichhorn et al., 2009; Perner et al., 2006; Saxe & Kanwisher, 2003).

In sum, tasks and stimuli that require, or robustly suggest, thinking about thoughts lead to activity in these regions. Thinking about thoughts can be manipulated by changing participants' instructions for the same stimuli, or by changing the stimuli with the same instructions. Very similar stimuli and tasks, however, which focus on physical objects, physical representations, or externally observable properties, do not lead to activity in these regions.

Links to behavior

The review in the previous section shows that tasks and stimuli that evoke thinking about thoughts also elicit metabolic activity in the ToM brain regions. However, there is even stronger evidence for a link between activity in these regions and understanding other minds: across trials, across individuals, and across development, performance on behavioral tests of ToM is related to brain activity in these same regions.

Most adults pass standard laboratory ToM tasks 100% of the time, leaving little room for inter-individual variability in accuracy. However, by using tasks with no simple right answer, it is possible to reveal individual differences in mental state attribution. Imagine, for example, learning about a girl Grace, who was on a tour of a chemical plant. While making coffee, Grace found a jar of white powder, labeled "sugar," next to the coffee machine. She put the white powder, which was actually a dangerous toxic poison, in someone else's coffee, who drank the poison and got sick. Is Grace morally blameworthy? These scenarios require weighing what Grace intended (her mental state) against what she did (the outcome), and participants disagree in their judgments; some people think she is completely innocent (because she believed the powder was sugar), whereas others assign some moral blame (because she hurt someone). This difference is correlated with neural activity during moral judgments across individuals; the more activity there was in a participant's right TPJ, in particular, the more the participant forgave Grace for her accidental harms (Young & Saxe, 2009b).

An alternative strategy is to measure the quantity and quality of people's spontaneous mentalistic attributions to ambiguous stimuli. For example, when viewing the simple animations of a small and large moving triangle (Castelli et al., 2000), people generate very rich mentalistic interpretations from the simple movements depicted in these stimuli (e.g. "the child is pretending to do nothing, to fool the parent"). People differ in the amount, and appropriateness, of the thoughts and feelings that they infer from the animations. People who have more activity in ToM brain regions, while watching the animations give more appropriate descriptions of the triangles' thoughts and feelings after the scan (Moriguchi et al., 2006). Relatedly, Wagner, Kelley, & Heatherton (2011) showed participants still photographs of natural scenes, approximately a quarter of which contained multiple people in a social interaction. Participants performed an orthogonal categorization task ("animal, vegetable, mineral?"). Individuals who scored high on a separate questionnaire, measuring tendencies to think about others' thoughts and feelings (the "empathizing quotient"; Baron-Cohen & Wheelwright 2004; Lawrence, Shaw, Baker, Baron-Cohen, & David, 2004) also showed higher activity in mPFC in response to photographs of interacting people.

Differences in ToM are easier to find in young children, who are still learning how to understand other minds. Interestingly, two recent studies from our laboratories suggest that getting older, and getting better at understanding other minds, is associated, overall, not with **more** activity in "ToM brain regions," but with more **selective** activity. Children from 5 to 12 years old all have adult-like neural activity when listening to stories about characters' thoughts and feelings. What is different is that ToM regions in younger children show similarly high activity when listening to any information about characters in the story, including the characters' physical appearance or social relationships (Saxe et al., 2009; Gweon et al., 2012), whereas in older children and adults, the ToM brain regions are recruited only when listening to information about thoughts and feelings (Saxe & Powell, 2006; Saxe et al., 2009). This developmental difference in the selectivity of the ToM brain regions is correlated with age, but also with performance outside the scanner on difficult ToM tasks (Gweon et al., 2012). Moreover, the correlation between neural "selectivity" and behavioral task performance remains significant in the right TPJ, even after accounting for age.

One limitation of all the foregoing studies is that they are necessarily correlational. The strongest evidence that some brain regions are involved in a cognitive task is to show that disrupting those regions leads to biases or disruption in task performance. Transcranial magnetic stimulation (TMS) offers a tool for temporarily disrupting a targeted brain region. We (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010a) compared people's moral judgments following TMS to either right TPJ or a control brain region 5 cm away. TMS to right TPJ, but not the control region, produced moral judgments temporarily biased away from considerations of mental state information. Innocent accidents appeared more blameworthy, while failed attempts appeared less blameworthy, as though it mattered less what the agent believed she was doing, and more what she actually did. People didn't lose the ability to make moral judgments altogether; they still judged that it is completely morally wrong to intentionally kill, and not wrong at all to simply serve someone soup. Disrupting the right TPJ thus appears to leave moral judgment overall intact, but impairs people's ability to integrate considerations of the character's thoughts into their moral judgments. Converging evidence comes from another TMS study: TMS to right TPJ made adults slower to recognize a false belief, in a simple (non-moral) false belief task (Costa, Torriero, & Oliveri, 2008).

Another way to study the necessary contributions of a brain region is to work with people who have suffered permanent focal (i.e. local) damage to that region, typically due to a stroke. Samson, Apperly, and colleagues (Apperly & Butterfill 2009; Apperly, Samson, & Humphreys, 2005; Samson, Apperly, & Humphreys, 2007) have conducted a series of elegant studies using this approach. Initially, these authors tested a large group of people, with damage to many different brain regions, on a set of carefully controlled tasks. They then identified individuals with a specific pattern of performance: individuals who passed all the control tasks (e.g. measuring memory, cognitive control, etc.), but still failed to predict a character's actions based on their false beliefs. Next, the scientists used a lesion-overlap analysis to ask which brain region was damaged in all, and only, the patients with this diagnostic profile of performance. The answer was the left TPJ, one of the same brain regions identified by fMRI.[3]

Summary

The literature from the last 10 years thus suggests a generalization—there are cortical regions in the human brain where activity is associated with **understanding other minds** in three ways:

1. Metabolic measures of activity reliably increase when the participant is thinking about thoughts, across a wide range of stimuli and tasks, but not in response to a variety of similar control tasks and stimuli.
2. Activity is correlated with behavioral measures of thinking about thoughts.
3. Disrupting activity leads to deficits in thinking about thoughts.

So far, these claims are relatively uncontroversial. As we noted above, there is a broad consensus in social cognitive neuroscience. However, much controversy remains about the proper **interpretation** of these data.

Exactly what is the nature of these regions, their functions, and their contribution to thinking about thoughts? Here's a strong hypothesis: one or more of these regions has the specific

[3] It is worth noting that the effects of lesions to the right TPJ, one of the regions argued to be most selective for ToM, haven't yet been effectively tested. The candidate participants all had extensive and diffuse damage to the right hemisphere, and failed the control tasks, making it impossible to test ToM deficits specifically.

cognitive function of representing people's mental states and experiences—that is, of thinking about thoughts. Whenever we are thinking about thoughts, there are neurons in these regions firing. These neurons are gathered in spatial proximity (i.e. into a "region") because they have related computational properties, that are distinct from the computation properties of neurons in the surrounding cortex.[4] The pattern of firing, in space and time, of these neurons encodes aspects of someone's thoughts. As an analogy, consider the way MT neurons encode speed and direction of motion, and face area neurons encode aspects of facial features that are relevant to face identity (Freiwald, Tsao, & Livingstone, 2009; Georgopoulos, Schwartz, & Kettner, 1986). In the proposed hypothesis, the neurons in the ToM brain regions encode aspects and dimensions of inferred thoughts. Scrambling the pattern of activity in these neurons would therefore lead to an inability to discriminate one inferred mental state from another, for example, making all minds appear homogenous: people might all seem to have the same desires and preferences, and the same knowledge and beliefs. More serious damage to these regions might make it impossible to think about other minds at all, without similarly impairing the rest of cognition.

There certainly is not enough evidence to prove that this strong hypothesis is right; a more immediate question is whether it is obviously wrong. There are at least two classes of potential objections: theoretical arguments, based on general principles of how the brain works, and empirical arguments, based on the results of other experiments in cognitive neuroscience. In the next section, we describe some of these objections, and some of our responses to them.

A strong hypothesis

Objections from theoretical considerations

Many authors have expressed discomfort with the project of trying to link specific cognitive functions with delineated brain regions. For example, a decade after their 2nd edition UoM chapter, Chris and Uta Frith wrote: "We passionately believe that social cognitive neuroscience needs to break away from a restrictive phrenology that links circumscribed brain regions to underspecified social processes" (Frith & Frith, 2012). Others have echoed this accusation of phrenology; for example, Bob Knight criticizes the "phrenological notion that a given innate mental faculty is based solely in just one part of the brain" (Knight, 2007), and William Uttal recently argued that "any studies using brain images that report single areas of activation exclusively associated with any particular cognitive process should a priori be considered to be artifacts of the arbitrary thresholds set by investigators and seriously questioned" (Uttal, 2011). Most such theoretical objections include variations on three themes: social cognitive neuroscientists are accused of (incorrectly) viewing regions as (1) functioning in isolation, (2) internally functionally homogenous, and (3) spatially bounded and distinct. Here, we address each of these concerns in turn.

First, does claiming that a region has a specific function (e.g. in thinking about thoughts) entail suggesting that this region functions in isolation? To put it more extremely, are we claiming that, for example, the right temporo-parietal junction (RTPJ) could pass a false belief task on its own? Obviously not. Performing any cognitive task necessarily depends on many different cognitive and computational processes, and therefore brain regions. No interesting behavioral task can be accomplished by a single region. The tasks of the mind and brain—recognizing a friend, understanding a sentence, deciding what to eat for dinner—must all be accomplished by a long sequence

[4] These distinct properties may derive entirely from patterns of connectivity, not from the structure of the neurons themselves.

of processing steps, passing information between many different regions or computations, from sensory processing all the way to motor action. The function of a neuron or a brain region should never be identified with completing a cognitive task. Thus, for example, "passing a false belief task" is not even a candidate function of a brain region. Any time an individual passes a false belief task, many brain regions—involved in perceiving the stimuli, manipulating ideas in working memory, making a decision, and producing a response—will all be required (Bloom & German, 2000).

More generally, we expect that the functions of regions (or neural population, regardless of spatial organization) will be to receive a class of inputs, and transform them into output, which make different information relatively explicit. Therefore, the specific questions about any neural population should include: what input does it receive, what output does it produce, and what information is made explicit in that transformation? Of course, the answers to these questions will require us to understand the position of this neural population within a larger network, especially when characterizing a region's input and output. In the case of the ToM regions, a related question concerns the relationships between the different regions within the network. At least five cortical regions are commonly recruited during many different social cognitive tasks: how is information passed between, and transformed by, each of these spatially distinct regions?

Thus, studying the function of a brain region means studying in isolation one component of a system that could never function in isolation. This description may sound ominous, but scientific progress frequently requires us to break complex systems into component parts. While the pieces could not function in isolation, understanding their isolated contributions is necessary to understanding the function of the integrated system. For any given neural population, it is reasonably to ask: what classes of stimuli and tasks predictably and systematically elicit increased activity in the population as a whole? Which dimensions of stimuli lead to activity in distinct subpopulations of neurons? Both traditional and new fMRI methods help answer these questions, albeit somewhat indirectly (more on this, in "Where next?").

The second objection is that studying brain regions leads to the false assumption that groups of spatially adjacent neurons are functionally homogenous. The regions we study in fMRI are orders of magnitude larger than what we believe are the true computational units of brain processing, the neurons. Changes in blood oxygenation measured by fMRI inevitably reflect averages over the activity of many thousands of individual neurons. Why is it useful to study oxygen flow to chunks of cortex approximately 1–5 cm^2 in size, which are so much bigger than neurons, but so much smaller than the networks required to complete a task?

Our suggestion is that there is no reason, a priori. It just happens, as a matter of empirical fact, that many interesting computational properties of the brain can be detected by studying the organization of neural responses on this scale. Aggregating the responses of neighboring neurons often produces informative population averages. Results obtained via fMRI reflect the same distinctions found in directly observable population codes (e.g. Kamitani and Tong, 2005; Kriegeskorte & Bandettini, 2007). This empirical fact may have a theoretical explanation. Neurons with similar or related functions may be spatially clustered to increase the computational efficiency of frequent comparisons (e.g. lateral inhibition). Blood-oxygen delivery to the cortex may follow the contours of neural computations, to increase the hemodynamic efficiency of simultaneously getting oxygen to all of the neurons that need it (Kanwisher, 2010). However, these arguments are not necessary premises; for fMRI to be useful, we only need the empirically observable fact that useful and reliable generalizations can be made for hemodynamic activity in patches of cortex at the spatial scale of millimeters.

Of course, though, neurons within a region or an fMRI voxel are never functionally homogenous. Consider the analogy of primary visual cortex—neurons in V1 have visual receptive fields, meaning that activity can be induced by a pattern of bars of light falling on a specific region of the retina. However, neurons in V1 differ from one another in where on the retina one should place the light (retinotopy), how large the pattern should be (size and spatial frequency preferences), and the orientation to which the bars should be rotated (orientation selectivity), to elicit a maximal response. There is also a separate (but systematically interleaved) population of neurons for which the response depends on color, but not orientation or size. Furthermore, some neurons primarily send information to subsequent regions of visual cortex (e.g. excitatory neurons) whereas other neurons primarily modulate the response of neighboring neurons in V1 (e.g. inhibitory interneurons). As far as we know, the metabolic activity measured by fMRI reflects a combination of activity in all of these different populations. Consequently, we should never assume that the amount of "activity" we measure in a region with fMRI represents (or would correlate very well with) the rate of firing of any individual neuron inside that region. Similarly, we cannot assume that if two different stimuli or tasks elicit similar magnitudes of activity in a region, then they are eliciting responses in the same, or even shared, neural populations. Completely non-overlapping subpopulations of neurons could produce the same magnitude of fMRI activity within a region. Any interpretation of fMRI data must be sensitive to this possibility. In fact, studying the organization of functional subpopulations within a region (e.g. which dimensions of stimuli are represented by distinct subpopulations within a region) may be one of the most powerful ways that fMRI will contribute to the neuroscience of ToM. We describe these methods in greater detail in "Where next?."

The third potential objection is that studying regions with fMRI leads researchers to imagine boundaries between discrete regions, when the truth is a continuous distribution of neural responses over the cortical sheet. The data we described in the first section shows that cortex is not functionally homogenous with regard to theory of mind, and regional distinctions are not all "artifacts of arbitrary thresholds." Still, there is a legitimate reason why cognitive neuroscientists may be reluctant to call any reliable functional regularity discovered by fMRI a "region." These "regions" may turn out to be just one piece of a larger continuous functional map over cortex, not computationally distinct areas of their own (Kriegeskorte, Goebel, & Bandettini, 2006).

Cortical responses at scales measurable by fMRI are organized along multiple orthogonal spatial principles. One is the division of cortex into cytoarchitectonic "areas," like primary visual cortex (V1) and primary auditory cortex (A1). Orthogonal to the division of cortex into areas are topographic principles. Most visual regions, for example, are organized by retinotopy: moving across the cortical sheet, the region of the visual field eliciting a maximal response varies smoothly, covering the whole visual field from fovea to periphery, top to bottom, and left to right. Likewise, multiple distinct motor areas are organized by somatotopy, and auditory areas by tonotopy.

These orthogonal principles of cortical organization create a challenge for cognitive neuroscientists, because in charting new territory, away from well-understood sensory and motor systems, we may claim to discover new functional regions associated with higher-order cognitive processes, which are really just one end of a larger map (cf. Konkle & Oliva, 2012). To make the puzzle concrete, imagine looking at functional responses to visual stimuli across occipital cortex for the first time, without the benefit of the history of visual neuroscience. One tempting way to divide the occipital cortex into functional "regions" might be by retinotopy—one group of patches that responds to foveal stimuli, and a different group of patches that responds to peripheral stimuli. This foveal vs. peripheral difference is highly robust, replicable within and across subjects, within and across tasks, and correlated with behavior (i.e. visual performance in corresponding regions of the visual field). Nevertheless, other considerations, such as cytoarchitecture, connectivity, and processing time, suggest that this is the wrong division for capturing functional and computational regularities. Identifying a robust functional regularity that divides one patch of cortex from another is not the same thing as identifying a true cortical area—a region that is computationally distinct from its neighbors, with distinct cytoarchitecture, connectivity, and topography (Friston, Holmes, Worsley, Poline, Frith, & Frackowiak, 1995; Kanwisher, 2010; Kriegeskorte et al., 2006; Worsley, Evans, Marrett, & Neelin, 1992). Instead of studying the "peripheral patches," neuroscientists divide the cortex into V1, V2, V3, MT, etc., each of which is then internally organized (in part) by retinotopy.

Imagine you are a social cognitive neuroscientist, looking at a new bit of cortex, and you see a new functional regularity—a patch of cortex that shows a high response when the individual is thinking about stories, cartoons, or movies depicting the contents of another mind. Which kind of functional regularity is this? On the one hand, these data could signal the discovery of a true computational area, like V1. On the other hand, the observed functional regularity might be more like "peripheral patches," one part of a stimulus space or dimension that is mapped across cortex, but crosscuts multiple computational areas.

We believe that current fMRI data cannot resolve this puzzle directly. One approach may therefore be to suspend judgment until other sources of evidence are available. If patches of cortex involved in understanding other minds are true cortical areas, it will be possible to distinguish them from their anatomical neighbors by cytoarchitecture, connectivity, and/or topography. Non-invasive functional imaging tools do not yet have high enough resolution to reveal cytoarchitecture in vivo. To study the links between function and cytoarchitecture with current technology, we would need to collect functional data to identify the regions *in vivo*, and then analyze the neuroanatomy of the same individual post mortem.

A weaker, but more accessible, source of evidence is patterns of connectivity. In some cases, cortical areas can be differentiated by their profiles of connectivity. It has become increasingly possible to measure the pattern of connections between regions using neuroimaging. Diffusion imaging (DTI), which looks at the predominant direction of water diffusion, allows us to visualize the dominant pathways of axons connecting brain regions. Using diffusion, we can ask whether a patch of cortex involved in understanding other minds shows a different pattern of connectivity than its neighbors to the rest of the brain. Initial evidence suggests it does, at least in the case of the right TPJ. Mars, Sallet, Schüffelgen, Jbabdi, Toni, & Rushworth (2012) found that the broad area of right temporo-parietal cortex (BA 39/40) can be sub-divided into three clusters, based on DTI connectivity alone, and one of these clusters is functionally correlated (during a resting baseline) with the other ToM regions, including medial prefrontal and medial parietal cortex. Although Mars and colleagues did not directly test whether the region defined by connectivity and the region defined by function (i.e. active in ToM tasks) share the same boundaries, the results are suggestive.

Currently, however, neither cytoarchitecture nor connectivity analyses give a definitive evidence that patches of cortex recruited by mental state reasoning tasks are true cortical areas. An alternative approach might therefore be to consider the alternative hypothesis directly—these regions are one part of a larger, continuous map. If we compare the cortical patches we are studying to their anatomical neighbors, is there a plausible higher-level "stimulus space," which could unite these responses into one map?

Again, the right TPJ is an interesting example. The right TPJ region that is activated by ToM tasks (as described in "Theory of mind and the brain") has two very close anatomical neighbors. One neighbor (up toward parietal cortex in the right inferior intraparietal sulcus (IPS), but confusingly sometimes also called RTPJ) is recruited by unexpected events that demand attention. These events may be unexpected because they are rare, or because a generally reliable cue was misleading on this occasion. Redirecting attention toward an unexpected event leads to metabolic activity in this region (Corbetta & Shulman, 2002; Mitchell, 2008; Serences, Shomstein, Leber, Golay, Egeth, & Yantis, 2005). Damage to this region makes it hard for objects and events in the contralateral visual field to attract attention, producing left hemifield neglect (Corbetta, Patel, & Shulman, 2008). Some authors have noted that false belief tasks typically involves an unexpected event (Corbetta et al., 2008; Decety & Lamm, 2007; Mitchell, 2008): for example, false belief tasks often hinge on an object unexpectedly changing location while the protagonist is out of room, and require the participant to shift attention between both locations. In fact, the region that is recruited during exogenous attention tasks is so close to the region recruited during ToM tasks that some concluded that these are actually just two different ways to identify the same region (Corbetta, Kincade, Ollinger, McAvoy & Shulman, 2000; Mitchell, 2008). More recently, however, both meta-analyses and high-resolution scanning within individual subjects suggest that there are actually two distinct cortical regions (or "patches"), and the region recruited by attention tasks is approximately 10 mm superior to the region recruited by ToM tasks (Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009; Decety & Lamm, 2007). Also, the region that is recruited during false belief tasks is not recruited by unexpected transfers of location in stories about false maps and physical representations, as described above (e.g. Young et al., 2010b). Nevertheless, it remains an interesting possibility that the exogenous attention response and the ToM response are part of a larger continuous map across cortex, a topography of different kinds of unexpected attentional shifts. The more superior end of this map could direct attention toward unexpected positions in space and time, while the more inferior end of the map could direct attention toward unexpected people, actions, or inferred thoughts.

A second topographical "stimulus space" of which the RTPJ could be a part runs anteriorly, through the right superior temporal sulcus (STS), and down toward the temporal pole. As with the RTPJ and the right IPS, the RTPJ and the posterior STS were initially conflated, but have subsequently been shown to be spatially distinct (Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007). Multiple parts of the STS are recruited when participants observe other people's actions in photographs, film clips, point light walkers or animations (Pelphry, Mitchell, McKeown, Goldstein, Allison, & McCarthy, 2003; Pelphrey, Morris, Michelich, Allison, & McCarthy, 2005; Hein & Knight, 2008). The STS is large, and Kevin Pelphrey and colleagues (Pelphrey et al., 2005; Pelphrey & Morris 2006) propose that it contains a pseudo-somatotopic map of observed actions: others' mouth motions represented most anteriorly, followed by hand and body movements, followed by head and eye movements represented most posteriorly. An intriguing possibility is that the temporo-parietal junction, which is at the most posterior end of the superior temporal sulcus, is part of this same map. Whereas hand and body movements convey information about what a person is **doing** and **intending**, head and eye movements convey information about what a person is **looking at** and **seeing**. Thus, the STS may contain a map of others actions that move from externally observable body movements (anterior end) toward invisible mental states (posterior end), culminating in the RTPJ, which responds to thinking about what a person is **thinking**.

In principle, either of these mapping hypotheses, or both, could be true. Evidence that a ToM region, e.g. the RTPJ, is part of a larger cortical map might come in the form of a response that moves continuously across contiguous patches of cortex, modulated by continuous changes in a stimulus dimension (Konkle & Oliva, 2012). For example, if the RTPJ is one end of an attention map, we might expect to see that surprising facts about physical entities elicit activity relatively far from the TPJ, but that as the unexpected stimulus becomes more social and interpersonal, activity moves continuous across the map, ending at the RTPJ. Similarly, if the RTPJ is the "abstract" or "head" end of a map of observable social actions, we'd expect to see continuous changes in the location of activity, as depictions of human actions become either more abstract or physically higher on the body. Finally, both could be true. The RTPJ could exist at its precise location because that is where a region that deals with unexpected information converges with a region that deals with human actions. We would be enthusiastic to see someone test these hypotheses further.

For now, however, we suggest that it does not matter for any of our empirical purposes whether the RTPJ (or any other patch of cortex recruited in ToM tasks) is a true cortical area, or whether it is one end of a larger map, as defined by cytoarchitecture, connectivity, and topography. In either case, there is a robust functional regularity replicable within and across subjects, within and across tasks, and correlated with behavior: a bounded, adjacent patch of cortex where activity is high during a range of ToM tasks. Without committing to whether they are true computational "areas," or just the equivalent of "Peripheral Patches," we believe that it is fruitful to continue to study these coherent patches as "regions," in order to discover the computations and representations that underlie thinking about other minds.

Given everything else we know about the brain, it is not surprising that systematic response profiles are linked to regions of cortex much bigger than a neuron and much smaller than a network. The challenge now is to integrate the huge, and growing, list of empirical discoveries and to construct hypotheses about the computations and representations in the neural populations we are studying. These empirical results form the basis of a different set of potential objections to the hypothesis that there is a strong specific link between cortical regions, like the TPJ, mPFC, and PC, and thinking about thoughts.

Empirical objections

An empirical objection to our argument in "Theory of mind and the brain" might begin by pointing out that our review of the literature was selective. In addition to the dozens of articles we cited, there are dozens of others that claim so-called "ToM regions" are active in tasks that do not involve thinking about thoughts, or are not active in tasks that do involve thinking about thoughts. How can we integrate these other data into a coherent hypothesis?

First, what about claims that "ToM regions" are active in tasks that do not involve thinking about thoughts? The RTPJ is again a useful example. As we mentioned above, some authors initially believed that the same region of RTPJ involved in false belief tasks was also recruited during any exogenous shift of attention and/or biological motion perception. Other literature suggests that the RTPJ is involved in maintaining a representation of one's own body, by integrating multi-sensory information and locating the body in space (Blanke et al., 2005; Blanke & Arzy, 2005; Tsakiris et al., 2008). Experiments that ask people to mentally rotate their own body or imagine their body in different parts of space, as well as those that induce changes in bodily self-perception (with e.g. rubber hands) find activation in this region (Arzy, Thut, Mohr, Michel, & Blanke, 2006; Blanke & Arzy, 2005). TMS and intracranial stimulation to this region has been shown to lead to out-of-body experiences, confusion of the body vs. the environment, and illusory changes in the orientation of body parts (Blanke et al., 2002, 2005; Tsakiris et al., 2008).

How do we interpret these results, which seem to contradict our theory? In general, we could consider five possibilities. The first is that one group of scientists has made an error, leaving an unnoticed confound in their experimental paradigm. Could the body representation tasks also involve thinking about thoughts? Could the false belief task accidentally induce updating maintaining a representation of one's own body? These are empirical questions, but we consider both possibilities highly unlikely. The second option is that there is some deep common computation, served by the same neural population that is required by these different classes of tasks. One option here would be "decoupling" (Gallagher & Frith, 2003; Leslie & Frith 1990; Liu, Sabbagh, & & Gehring, 2004), maintaining distinct representational reference frames, e.g. for one's own and other minds, for current vs. imaginary body positions, and so on. The third option is that the same neurons can assume distinct and even unrelated functional roles, depending on the context and the pattern of activity in other neural populations (e.g. Miller & Cohen, 2001). On this view, thinking about thoughts and maintaining a body representation are cognitively unrelated, in spite of being implemented by the same neurons. The fourth option is that distinct neuronal populations are involved in thinking about thoughts and maintaining a body representation, but these neurons are interleaved within the RTPJ, just as color- and orientation-sensitive neural populations are interleaved in V1.

Finally, the fifth option is that distinct neuronal populations are involved in thinking about thoughts vs. maintaining one's body representation, exogenous attention shifts, and biological motion perception. These neural populations are not interleaved: they are contained in distinct regions that are merely nearby on the cortical sheet. We find this last option most plausible. Standard fMRI methods, which involve extensive spatial blurring at three stages (acquisition, preprocessing, and group averaging; Fedorenko & Kanwisher, 2009; Logothetis, 2008), are strongly biased to conflate neighboring regions that are truly distinct. Even so, the existing evidence suggests that the region involved in body representations is lateral (MNI x-coordinates typically around 64mm) to the region involved in thinking about thoughts (x-coordinates around 52 mm). As described above, this was also true of the regions involved in biological motion perception (Gobbini et al., 2007) and exogenous attention (Decety & Lamm, 2007; Scholz et al., 2009); these almost completely non-overlapping in individual subjects.

A different kind of challenge arises from examples of tasks that apparently do involve thinking about thoughts, but do not elicit activity in TPJ, mPFC, or PC. Two interesting examples are visual perspective-taking tasks, and recognition of facial expressions of emotions from photographs. In visual perspective taking tasks, the participant sees an image of a character in a 3D space, and is asked to imagine the view of the room from the viewpoint of the character. For example, participants may be asked to report how many dots the character can see (the third person perspective), versus how many dots the participants themselves can see (including those that are out of the character's view, the first person perspective). This perspective-taking task clearly involves thinking about the character's visual access, which could be construed as a mental state. Nevertheless, this task typically does not elicit activity in the same regions as thinking about thoughts (Aichhorn et al., 2006; Vogeley May, Ritzl, Falkai, Zilles, & Fink, 2004). Similarly, participants in hundreds of fMRI experiments have viewed photographs of human faces expressing various basic emotional expressions (e.g. sad, afraid, angry, surprised, happy, neutral). Although these images do depict evidence of another person's emotional experience, they also typically do not elicit activity in the same regions as thinking about thoughts (Costafreda et al., 2008; Lamm, Batson, & Decety, 2007; Vuilleumier et al., 2001).

What should we conclude from these examples? One option is to make a forward inference. Using the evidence from these tasks to change our hypothesis about the brain regions' functions. Here, the forward inference might be that the ToM brain regions are responsible for only a subset of mental state processing. We could conclude that different brain regions are involved in thinking about different classes of internal experiences: bodily states, emotional states, perceptual states, or epistemic states (e.g. thinking, knowing, doubting, etc.). The so-called ToM brain regions might be specifically involved in representing epistemic states, while regions of insula represent others' emotions and regions of parietal cortex represent others' perceptual states.

Another option is to make a reverse inference: using the pattern of neural activity to change our analysis of the cognitive processes required by the task. Reverse inferences are risky because they require a lot of confidence in the functional specificity of the brain region(s) involved (Poldrack, 2006a). However, we think ToM brain regions are good candidates to support reverse inference, given the converging evidence across the many experiments described above. In this case, a reverse inference might be that these visual perspective taking and emotional face tasks do not actually elicit thinking about thoughts, and instead are solved by alternative computational strategies. For example, rather than truly considering someone's perceptual experiences, line-of-sight tasks may be solved using mental rotation and geometric calculation. Emotional facial expressions may be recognized (akin to object recognition) without always requiring a representation of the person's internal state.

In these particular cases, we are open to either the forward or reverse inference; only further experimentation will tell which is the better generalization. In general, we believe that in this domain, inferences can be made in both directions, forward and reverse. The absence of activation in perspective taking and emotion recognition tasks provides an important constraint on the possible functions of ToM brain regions (the forward inference). At the same time, the fact that visual perspective-taking and emotion recognition rely on different brain regions from reading stories about thoughts provides evidence that these tasks depend on different cognitive processes (the reverse inference). It's the give and take of these two kinds of inferences, as evidence accumulates, that allows us to build a coherent understanding of both cognitive and neural function.

Summary

Taken together, these first two sections illustrate the main contributions of the first decade of neuroimaging the understanding other minds. We have discovered a robust, replicable functional regularity in the human brain: regions that have increased activity when participants think about thoughts. These regions may be true cortical areas or parts of larger topographical maps, but in either case, understanding other minds is a major organizing principle of responses over cortex. The function of these regions is not to complete a task, but to transform some class of input into some output; and the class of input has something to do with thinking about minds, and not bodies or abstract representations. Of course, this description remains unsatisfying and largely underspecified. Nevertheless, we are optimistic that the second decade of this research program will continue to improve our specifications. In particular, we are excited about newly emerging methods for fMRI data analysis that focus on the second aspect of a region's function: the features within its preferred stimulus class that organize differential responses within each of the ToM regions.

Where next?

Differences between theory of mind regions

One step in specifying the computations performed by ToM regions will be understanding the division of labor and information transfer between the different regions. Overall, ToM regions show similar profiles to most of the contrasts we described. However, research in the last 5 years has begun to tease apart the functional profiles of these regions, and the differences are intriguing, though much work still remains to be done to form a coherent view of how they all fit together.

The most striking contrast comes from a task that elicits very robust activity in the medial ToM regions (mPFC and precuneus), but not the lateral ToM regions (TPJ and anterior STS): thinking about personality traits, especially of the self and close others (Whitfield-Gabrieli et al., 2011; Moran et al., 2011a; Saxe, Moran, Scholz, & Gabrieli, 2006a; Krienen, Tu, & Buckner, 2010). In a typical version of the experiment, participants in the scanner see single words describing personality traits (e.g. "lazy," "talkative", "ambitious") and judge either whether each one is desirable or undesirable (the semantic control condition), is true of a famous person (the other control condition), or is true of themselves (the self condition). MPFC and precuneus regions show much higher activity during the self condition; moreover, within the self condition, activity in mPFC is higher for the words that participants say **are** true of themselves (Moran et al., 2011a), and the amount of activity in mPFC for each item presented in the self task (but not in the semantic or other tasks) predicts participants' subsequent memory for those items on a surprise memory test (Mitchell, Macrae, & Banaji, 2006; Jenkins & Mitchell, 2009). These data are compelling to us: the mPFC, but not the TPJ, is involved in reflection about one's own stable traits and attributes (Lombardo et al., 2010). Similarly, elaborating one's own autobiographical memories leads to activity in medial ToM regions, whereas imagining someone else's experiences on similar occasions elicits activity in bilateral TPJ (Rabin, Gilboa, Stuss, Mar, & Rosenbaum, 2010).

In fact, coding information in terms of similarity to the self may be a key computation of mPFC. In one series of studies (Tamir & Mitchell, 2010), participants judged the likely preferences of strangers (e.g. is this person likely to "fear speaking in public" or "enjoy winter sports") about whom they had almost no background information. Under those circumstances, the response of the mPFC (but not TPJ) was predicted by the discrepancy between the attributions made to the target and the participant's own preference for the same items: the more another person was perceived as different from the self, for a specific item, the larger the response in mPFC.

Another distinction, supported by multiple studies, suggests that sub-regions of mPFC are most recruited when thinking about someone's negative emotions or bad intentions, whereas the TPJ makes no distinction based on valence. For example, Bruneau, Pluta, & Saxe (2011) found that only the mPFC showed a higher response to stories about very sad events (e.g. a person proposes marriage and is rejected) compared to neutral or positive events (e.g. the marriage proposal is accepted). In a PET study, Hayashi, Abe, Ueno, Shigemune, Mori, Tashiro, et al. (2010) found that a region in mPFC was recruited when people were considering an actor's dishonesty as a factor in moral judgments; and in an fMRI study, Young & Saxe (2009a) found that a region in ventral mPFC was correlated with moral judgments of **attempted** harms, which are morally wrong only because of the actor's negative intentions. Converging with these neuroimaging studies, lesion studies suggest that focal damage to ventral mPFC creates disproportionate difficulty in understanding bad intentions, and in integrating those intentions into moral judgments (Koenigs et al., 2007).

In this vein, further work has been done to examine the response profile of mPFC. The "regions" implicated in ToM are very large, especially in the mPFC, and there may be multiple sub-divisions, each with different response profiles. Proposed distinctions along the ventral-dorsal axis of mPFC include: similarity to self (such that self-relevant processes elicit responses more ventrally (e.g. Mitchell et al., 2006; Jacques. Conway, Lowder, & Cabeza, 2011), interpersonal closeness (people who are closer, or more important to the self elicit responses more ventrally, Krienen et al., 2010), or affective content ("hot" affective states elicit responses more ventrally (Ames, Jenkins, Banaji, & Mitchell, 2008; D'Argembeau et al., 2007; Mitchell & Banaji, 2005), while "cool" cognitive states elicit responses more dorsally (Kalbe, Schlegel, Sack, Nowak, Dafotakis, Bangard, et al., 2010; Shamay-Tsoory & Aharon-Peretz, 2007; Shamay-Tsoory, Tomer, Berger, Goldsher, & Aharon-Peretz, 2005). Thus, while there may be a convergent theoretical account of the mPFC, and its responses to other people's negative intentions and emotions, one's own personality traits, and ambiguous inferences about preferences, another possibility is that these response profiles reflect distinct sub-regions within mPFC, each contributing a distinct computation to understanding other minds.

Intriguingly, none of these distinctions have shown to affect the lateral ToM regions. In contrast, one dimension that seems to influence the magnitude of response in TPJ more than mPFC is whether someone's thought or feeling is unexpected, *given the other information you have about that person*. Saxe & Wexler (2005) introduced characters whose social background was mundane (e.g. New Jersey) or unusual (e.g. a polyamorous cult). Participants then read about that character's thoughts and feelings (e.g. a husband who believed it would be either fun or awful if his wife had an affair). On their own, neither the background nor the content of the belief affected the magnitude of response in the TPJ, but there was a significant interaction: whichever thought was unlikely, given the character's social background, elicited a larger response in the right TPJ. Recently, Cloutier, Gabrieli, O'Young, & Ambady (2011) provided a conceptual replication of this result: participants saw photographs of people labeled as Democratic or Republican, paired with opinions that were either typical of their political affiliation or typical of the opposite affiliation. Opinions that were unexpected given the protagonist's political background (e.g. a Republican wanting liberal Supreme Court judges) elicited a higher response in most of the ToM regions, including bilateral TPJ and mPFC. Finally, a third study suggests that the conflict between background and belief is necessary for increased activation, not just sufficient. In the absence of specific background information about the believer, there is no difference in the response of any ToM region to absurd vs. commonsense beliefs (e.g. "John believes that swimming in the pool is a good way to grow fins/cool off," Young et al., 2010b).

Although they are preliminary, we find these results exciting because they are consistent with the idea that activity in TPJ reflects a process of forming a coherent model of another's mind. We expect other people to be coherent, unified entities, and strive to resolve inconsistencies with that expectation (see Hamilton & Sherman, 1996). Consequently, when a target's behavior violates a previous impression of that person, observers spend more time processing the behavior (Bargh & Thein, 1985; Higgins & Bargh, 1987) and more time searching for the cause of the behavior (Hamilton, 1988). Concomitantly, more activity in TPJ occurs precisely when participants are likely exerting effort to integrate a person's thoughts and feelings into a coherent model of their whole mind; that is, when participants are building a "theory" of a mind (Gopnik & Meltzoff, 1997).

These results suggest that while TPJ activity may be related to the discrepancy between a thought or feeling and other information about the protagonist, mPFC and PC activity may be related to discrepancies between the protagonist's thoughts or feelings and the participant's own thoughts or feelings on the same topic. Thus, whereas TPJ may be involved in integrating a belief or preference into a coherent model of another's mind (e.g. Young et al., 2010c), mPFC and precuneus activity may reflect a different "anchor-and-adjust" strategy, that helps identify other people's thoughts and preferences by starting with one's own preferences, and then adjusting them as necessary (Tamir & Mitchell, 2010). Taken together, these results suggest that medial and lateral ToM regions support distinct computations within ToM. These distinctions help us separate ToM into its real (i.e. neurally-realized) component parts, and formulate hypotheses about each of the more specific functions of individual regions within the group.

Magnitude: "more" theory of mind

Until now, we asked simply whether ToM brain regions do or do not show activity in response to a task or stimulus. However, this is clearly an over-simplification; activation is continuous, not discrete, making it very tempting to ask, "Which stimulus and task dimensions within the domain of thinking about minds modulate the activation in these brain regions?" The magnitude of activity, over tasks or stimuli, could reveal not only what class of stimuli is processed in a region, but also **which dimensions** of those stimuli and tasks elicit more or less processing.

Interestingly, initial attempts to modulate activation in the ToM network mostly discovered features that do not elicit differential magnitudes of response. For example, most of the ToM regions show an equally high response to explicit descriptions of beliefs that are true or false (Young & Saxe, 2008), justified or unjustified (Young, Nichols, & Saxe, 2010c), and well-intentioned or bad (i.e. a girl who believes she is putting poison in her friends coffee, vs. believes that she is putting sugar in the coffee, Young, Cushman, Hauser, & Saxe, 2007). In the TPJ, at least, it also does not matter to *whom* the thought or feeling is attributed: there is an equally strong response to beliefs attributed to similar or dissimilar others (Saxe & Wexler, 2005), or to members of one's own group vs. an enemy group (Bruneau & Saxe, 2010; Bruneau, Dufour & Saxe, 2012).

Part of the reason for this lack of success may be that we do not yet have satisfactory cognitive or computational theories of ToM that allow us to predict when "more" ToM processing will be required, or even exactly what "more" means. Some intuitive possibilities have already proven empirically false. For example, making it harder to infer what a character believes, by making the available evidence more ambiguous, does not lead to more activity in TPJ regions (Jenkins & Mitchell, 2009). In fact, we (Dodell-Feder et al., 2011) found that, while some stories about thoughts systematically elicit more activity than others, in each ToM region, we could not find any feature (e.g. vividness, unexpectedness, length, syntactic complexity) that predicted these differences in activity, with the partial exception of the precuneus, which showed greater activity to stories that involved more people.

Researchers with a background in computer science or game theory often suggest one particular dimension for "more" ToM processing—the depth of embedding of one mental state within another. Thus, many people intuit, reasoning about an embedded belief (e.g. "Carla believes that Ben thinks that she eats too much junk food") should require **more** ToM processing that reasoning about a simple belief (e.g. "Carla believes that she eats too much junk food.") When we tested this hypothesis directly using verbal stories, we found that no ToM regions showed greater activity for the more embedded beliefs (Koster-Hale & Saxe, 2011). Other brain regions did show more activity—regions involved in language processing and regions involved difficult memory and cognitive control tasks, like dorsolateral prefrontal cortex (DLPFC). Our interpretation of these results is that embedding beliefs inside other beliefs makes the reasoning problem harder, but does not lead to greater ToM processing, per se. This finding converges with results from patient populations and aging adults showing that failure to pass second-order false belief tasks may, in fact, be due to domain-general impairment, rather than diminished ToM processing (Slessor, Phillips, & Bull, 2007; Zaitchik et al., 2006).

This highlights a more general problem with trying to elicit ToM in games. Difficult games often demand ToM reasoning, but similar patterns of behavior can be achieved by logical problem solving. Thus, in games designed to allow for more or less sophisticated ToM reasoning, some papers find ToM regions correlated with increasing "levels of embedding" (e.g. Coricelli & Nagel, 2009), whereas other papers implicate control/memory brain regions, such as DLPFC (e.g. Yoshida et al., 2010). Some participants may, some of the time, discover non-mentalistic strategies to play the game, and patterns of play alone are less diagnostic than one might hope.

Thus, making progress in understanding what features or dimensions drive these regions, we believe, will again require both forward and reverse inferences. Cognitive or computational theories should suggest possible dimensions of ToM inferences that may be reflected in "more" activity in ToM regions; but at the same time, the dimensions that do, and do not, modulate the magnitude of response in ToM brain regions may provide important clues for developing theories of what these brain regions are actually doing.

Patterns within theory of mind regions

Finally, as well as looking at changes in overall activity, a third strategy is to look for divisions of functional responses across the neural populations within each region. Two relatively novel methods for analyzing fMRI data may allow neuroscientists to look inside ToM regions. Distinctions between neural subpopulations within regions may provide clues to how these regions function.

The first method is repetition-suppression, also called functional adaptation. Repetition-suppression analyses take advantage of the observation that after processing a stimulus or task once, activity in a neuron or brain region in response to an identical stimulus or task is suppressed, or adapted (Grill-Spector, Henson, & Martin, 2006). By manipulating the features of the repeated stimulus, so that some are identical and some are different from the original stimulus, it is possible to ask what **counts** as the same stimulus for a particular brain region (Kourtzi & Kanwisher, 2001). If the repeated stimulus is effectively the same with regard to the features represented by the brain region, then the region's response will be suppressed or adapted. On the other hand, if a feature that is represented in the brain region has been sufficiently modified, the brain region's response will "recover" from adaptation.

This method has been used frequently and effectively to study visual representations of objects and places (Poldrack, 2006b). To date, only one study has taken advantage of this approach to test hypotheses about ToM. Jenkins, Macrae, & Mitchell (2008) asked whether attributing a preference (e.g. "enjoys winter sports") to oneself, a similar stranger, or a dissimilar stranger, depend on the same neural subpopulations within mPFC. They found that thinking about a similar other person after thinking about the self led to repetition suppression, while thinking about the self and then a dissimilar other led to recovery from adaption. These results support the hypothesis that the mPFC represents other minds in terms of their similarity to the self.

Currently, we are using a similar strategy to investigate the components of mental state attributions. People read short stories in which a key mental state was repeated twice, with some elements changed. After the first mental state sentence (e.g. "Megan thinks that Julie is being too flirty"), the repetition either changed the agent (e.g. "**Gina** thinks"), the attitude verb (e.g. "Megan **worries** that"), the content (e.g. "Megan thinks that Julie **should be more flirty"**), all three, or none of these elements. In preliminary data, we find that ToM brain regions recover from adaptation for each kind of change on its own (compared to no change), suggesting that these regions encode all three elements of a mental state attribution. Interestingly, while the mPFC shows the most recovery when the content of the mental state changes, the left temporoparietal junction (LTPJ) shows the most recovery when the attitude verb changes, and the RTPJ shows equal recovery for any kind of change. If these results hold up to further analyses, they may provide clues about the contributions of each ToM brain region to thinking about the minds of others.

The second method, multi-voxel pattern analysis (MVPA), looks for subtle, but reliable spatial patterns within a single region (or local neighborhood) of cortex. By looking at spatial separability, MVPA provides a more direct measure of the existence of functionally separable sub-populations of neurons than repetition suppression. Each "voxel" (the fMRI equivalent of a pixel) may have some, possibly very small, preference for one kind of stimulus over another due to biases in the neuronal populations toward one type of information-processing vs. another; pattern analyses measure the similarities and differences between these patterns across space (Haxby, 2001). A few studies have begun to use pattern analyses to study ToM (e.g. Gilbert et al., 2008). In one promising example, Peelen, Wiggett, & Downing (2006) found that the pattern of activation in the mPFC reliably reflected the content of another person's emotion (e.g. sad vs. angry), independent of the stimulus modality (e.g. vocal expressions vs. body posture). Recently, we found that MVPA can be used to distinguish types of mental states within the RTPJ (Koster-Hale et al., 2013). Specifically, we find that the spatial patterns of responses across voxels (but not the magnitude of response) distinguished between harms committed intentionally vs. accidentally. This distinction cannot be detected in the pattern of activity in any other ToM brain region (or in any other part of the brain).

Moreover, we find that individual differences in the neural pattern predict individual differences in moral judgment: the individuals who have the most distinct neural patterns are also those who show the greatest behavioral difference in their moral judgments of intentional vs. accidental harms. Together, these results begin to show which distinctions are represented within a region, and point toward the underlying distinctions and computations in ToM. We are very excited by this line of inquiry, and expect that both repetition suppression and multi-voxel pattern analyses will be important contributors to the next decade of neuroimaging studies of ToM.

Limits of neuroimaging

We are optimistic that there is a lot still to learn about ToM from neuroimaging. We hope that the "neuroimaging" chapter of *Understanding Other Minds,* 4th edition will be as different from this one as this one is from the Friths' chapter in the 2nd edition. However, it is also important to be realistic. Neuroimaging is cumbersome, expensive, and fundamentally limited. Many basic questions about ToM cannot be addressed with neuroimaging. For example, a scientific theory of how humans understand other minds should address questions like: "When and why do we (spontaneously) seek to understand another's thoughts?", "How do we figure out the actual content of someone else's thoughts (i.e. *what* they are thinking) from specific cues?", "How do we choose whether or not to incorporate others' thoughts into our own decisions and actions?", and "Why do we care emotionally about others' thoughts and feelings?" None of these questions have yet been approached using neuroimaging, and may pose much harder challenges than the simpler questions we have addressed so far. Contemporary neuroimaging technology does not even allow us to address many fundamental questions about the neural mechanisms of ToM. Existing tools are extremely slow and blurry, by comparison to the speed and precision of neural computation: they cannot decipher what is the input of a region, how that input is transformed, or where the output from that region is sent, during a ToM task.

If our final horizon is a complete theory of how brain regions allow us to understand other minds, we will need to make dramatic progress on (1) the "psychophysics" of ToM in adulthood, to allow precise quantitative measurements of people's use of ToM; (2) a computational model of ToM that is sufficiently explicit to make quantitative predictions about adult judgments (e.g. Baker, Saxe, & Tenenbaum, 2011); and (3) a mechanism of how neurons and networks of neurons might implement that computational model, by sequentially transforming patterns of input into patterns of output that make different information explicit. That horizon is still far away. However, the fact that we can give a characterization of some of the boundary conditions that a successful account of ToM needs to meet is part of what makes this such an exciting time to participate in the cognitive neuroscience of understanding other minds.

References

Adams, R., Jr., Rule, N. O., Franklin Jr, R. G., Wang, E., Stevenson, M. T., Yoshikawa, S. (2010). Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience* **22**(1): 97–108.

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annual review of psychology* **60**: 693.

Adolphs, R., 2010. Conceptual challenges and directions for social neuroscience. *Neuron*, **65**(6): 752–67.

Aichhorn, M. et al. (2006). Do visual perspective tasks need theory of mind? *NeuroImage*, **30**(3): 1059–68.

Aichhorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: Falseness, beliefs, or attention. *Journal of Cognitive Neuroscience* **21**(6): 1179–92.

Ames, D. L., Jenkins, A. C., Banaji, M. R., & Mitchell, J. P. (2008). Taking another person's perspective increases self-referential neural processing. *Psychological Science* **19**(7): 642–4.

Apperly, I. & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states?. *Psychological Review* **116**(4): 953–70.

Apperly, I. A., Samson, D., Chiavarino, C., Bickerton, W. L., & Humphreys, G. W. (2007). Testing the domain-specificity of a theory of mind deficit in brain-injured patients: Evidence for consistent performance on non-verbal, "reality-unknown" false belief and false photograph tasks. *Cognition* **103**(2): 300–321.

Apperly, I. A., Samson, D., & Humphreys, G. W. (2005). Domain-specificity and theory of mind: evaluating neuropsychological evidence. *Trends in Cognitive Sciences* **9**(12): 572–7.

Arzy, S., Thut, G., Mohr, C., Michel, C. M., & Blanke, O. (2006). Neural basis of embodiment: distinct contributions of temporoparietal junction and extrastriate body area. *Journal of Neuroscience* **26**(31): 8074–81.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: modeling joint belief-desire attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society,* pp. 2469–74.

Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: The case of information overload. *Journal of Personality and Social Psychology* **49**(5): 1129.

Baron-Cohen, S., O'Riordan, M., Jones, R., Stone, V., & Plaisted, K. (1999). A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome. *Journal of Autism and Developmental Disorders* ***29***, 407–18.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders* **34**(2): 163–75.

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): 241–51.

Bedny, M., Pascual-Leone, A., & Saxe, R. R. (2009). Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences of the United States of America* **106**(27): 11312–17.

Bhatt, M., Lohrenz, T., Camerer, C. F., & Montague, P. R. (2010). Neural signatures of strategic types in a two-person bargaining game. In *Proceedings of the National Academy of Sciences* **107**(46): 19720–5

Blanke, O., & Arzy, S. (2005). The out-of-body experience: disturbed self-processing at the temporo-parietal junction. *Neuroscientist* **11**(1): 16–24.

Blanke, O., Mohr, C., Michel, C. M., Pascual-Leone, A., Brugger, P., Seeck, M., et al. (2005). Linking out-of-body experience and self processing to mental own-body imagery at the temporoparietal junction. *Journal of Neuroscience* **25**(3): 550–7.

Bloom, P., & German, T. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* **77**(1): B25–31.

Bruneau E., Dufour N., & Saxe R. (2012) Social cognition in members of conflict groups: behavioural and neural responses in Arabs, Israelis and South Americans to each other's misfortunes. Philosophical Transactions of the Royal Society, London, B Biological Sciences **367**(1589): 717–30.

Bruneau, E. G., Pluta, A., & Saxe, R. (2011). Distinct roles of the "shared pain" and "theory of mind" networks in processing others" emotional suffering. *Neuropsychologia* pp.1–13.

Bruneau, E. G., & Saxe, R. (2010). Attitudes toward the outgroup are predicted by activity in the precuneus in Arabs and Israelis. *NeuroImage* **52**(4): 1704–11.

Carrington, S. J., & Bailey, A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping* **30**(8): 2313–35.

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): 314–25.

Cloutier, J., Gabrieli, J. D. E., O'Young, D., & Ambady, N. (2011). An fMRI study of violations of social expectations: When people are not who we expect them to be. *NeuroImage* **57**(2): 583–8.

Cohen, S. B., Wheelwright, S., & Hill, J. (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): 241–51.

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience* **3**: 292–7.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Review Neuroscience* **3:** 201–15.

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

Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences* **106**(23): 9163–8.

Costa, A., Torriero, S., & Oliveri, M. (2008). Prefrontal and temporo-parietal involvement in taking others' perspective: TMS evidence. *Behavioural Neurology* **19**(1–2): 71–4.

Costafreda, S. G., Brammer, M. J., David, A. S., & Fu, C. H. (2008). Predictors of amygdala activation during the processing of emotional stimuli: A meta-analysis of 385 PET and fMRI studies. *Brain Research Reviews* **58**(1): 57–70.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* **108**(2): 353–80.

D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., et al. (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.

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): 580–93.

Decety, J., Michalska, K. J., & Akitsuki, Y. (2008). Who caused the pain? An fMRI investigation of empathy and intentionality in children. *Neuropsychologia* **46**(11): 2607–14.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage* **55**(2): 705–12.

Döhnel, K., Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., & Sommer, M. (2012). Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *NeuroImage* **60**(3): 1652.

Dufou,r N., Redcay, E, Young, L., Mavros, P., Moran, J., Triantafyllou, C., Gabrieli, J., & Saxe, R. (2012). What explains variability in brain regions associated with Theory of Mind in a large sample of neurotypical adults and adults with ASD? In N. Miyake, D. Peebles, & R. P. Cooper (Eds), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 312–17). Austin: Cognitive Science Society.

Fedorenko, E., & Kanwisher, N. (2009). Neuroimaging of language: Why hasn't a clearer picture emerged? *Language and Linguistics Compass* **3**(4): 839–65.

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

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature* **12**(9): 1187–96.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping* **2**: 189–210.

Frith, C. D., & Frith, U. (2000). The physiological basis of theory of mind. In: S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds), *Understanding Other Minds: Perspective from Developmental Social Neuroscience* (pp. 335–56). Oxford: Oxford University Press.

Frith, C., & Frith, U. (2012). Social neuroscience. *Annual Review of Psychology* **63**: 287–313.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of "theory of mind". *Trends in Cognitive Sciences* **7**, 77–83.

Gallagher, H., 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.

Georgopoulos, A., Schwartz, A., & Kettner, R. (1986). Neuronal population coding of movement direction. *Science* **233**(4771): 1416–19.

Gilbert, S. J., Meuwese, J. D., Towgood, K. J., Frith, C. D., & Burgess, P. W. (2009). Abnormal functional specialization within medial prefrontal cortex in high-functioning autism: a multi-voxel similarity analysis. *Brain* **132**(4): 869–78.

Gobbini, M., 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): 1803–14.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *NeuroReport* **6**: 1741–6.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories.* Cambridge, MA*:* MIT Press.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences* **10**(1): 14–23.

Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child Development* **83**: 1853–68.

Hamilton, D. L. (1988). Causal attribution viewed from an information-processing perspective. In: D. Bar-Tal & A. W. Kruglanski (Eds), *The Social Psychology of Knowledge* (pp. 359–85). Cambridge: Cambridge University Press.

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review* **103**(2): 336.

Haxby, J. V. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**(5539): 2425–30.

Hayashi, A., Abe, N., Ueno, A., Shigemune, Y., Mori, E., Tashiro, M., & Fujii, T. (2010). Neural correlates of forgiveness for moral transgressions involving deception. *Brain Research* **1332**: 90–9.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology* **57**(2): 243–59.

Hein, G., & Knight, R. T. (2008). Superior temporal sulcus-it's my area: or is it? *Journal of Cognitive Neuroscience* **20**(12): 2125–36.

Higgins, E. T., & Bargh, J. A. (1987). Social cognition and social perception. *Annual Review of Psychology* **38**(1): 369–425.

Jacques, P., Conway, M. A., Lowder, M. W., & Cabeza, R. (2011). Watching my mind unfold versus yours: An fMRI study using a novel camera technology to examine neural differences in self-projection of self vs. other perspectives. *Journal of Cognitive Neuroscience* **23**(6): 1275–84

Jenkins, A., Macrae, C., & Mitchell, J. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences* **105**(11): 4507.

Jenkins, A., & Mitchell, J. (2010). Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex* **20**(2): 404–410.

Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., et al. (2010). Dissociating cognitive from affective theory of mind: a TMS study. *Cortex*, **46**(6): 769–80.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* **8**(5): 679–85.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. In *Proceedings of the National Academy of Sciences* **107**(25): 11163–70.

Kobayashi, C., Glover, G. H., & Temple, E. (2007). Children's and adults'neural bases of verbal and nonverbal "theory of mind". *Neuropsychologia* **45**: 1522–32.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* **446**(7138): 908–11.

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipto-temporal cortex. *Neuron* **74**(6): 1114–24.

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* **110**(14): 5648–5653.

Koster-Hale, J., & Saxe, R. R. (2011). Theory of Mind brain regions are sensitive to the content, not the structural complexity, of belief attributions. In: L. Carlson, C. Hoelscher, & T. F. Shipley (Eds), *Proceedings of the 33rd Annual Cognitive Science Society Conference,* pp. 3356–61.

Kourtzi, Z. & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science* **293**(5534): 1506.

Knight, R. T. (2007). Neural networks debunk phrenology. *Science* **316**(5831): 1578–9.

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage* **38**(4): 649–62.

Kriegeskorte, N., Goebel, R., & Bandettini, P. A. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences USA* **103**(10), 3863–8.

Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience* **30**(41): 13906–15.

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience* **19**(1): 42–58.

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004). Measuring empathy: reliability and validity of the empathy quotient. *Psychological Medicine*, **34**(5): 911–24.

Leslie, A. M., & Frith, U. (1990). Prospects for a cognitive neuropsychology of autism: Hobson's choice. *Psychological Review* **97**(1): 122–31.

Liu, D., Sabbagh, M., A., Gehring, W. J., & Wellman, H. M. (2004). Decoupling beliefs from reality in the brain: an ERP study of theory of mind. *Neuroreport* **15**(6): 991–5.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, **453**(7197): 869–78.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., et al. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience*, **22**(7): 1623–35.

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

Mason, R. A., & Just, M. A. (2010). Differentiable cortical networks for inferences concerning people's intentions vs. physical causality. *Human Brain Mapping*, **32**(2): 313–29.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**: 167–202.

Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, **18**(2): 262–71.

Mitchell, J. P., Banaji, M. R., & MacRae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience* **17**(8): 1306–15.

Mitchell, J., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, **50**(4): 655–63.

Moran, J. M., Lee, S. M., & Gabrieli, J. D. E. (2011a). Dissociable neural systems supporting knowledge about human character and appearance in ourselves and others. *Journal of Cognitive Neuroscience*, **23**(9): 2222–30.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011b). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, **108**(7): 2688–92.

Moriguchi, Y., Ohnishi, T., Lane, R. D., Maeda, M., Mori, T., Nemoto, K., et al. (2006). Impaired self-awareness and theory of mind: An fMRI study of mentalizing in alexithymia. *NeuroImage* **32**(3): 1472–82.

Onishi, K., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, **308**(5719): 255.

Peelen, M. V., Wiggett, A. J., & Downing, P. E. (2006). Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron*, **49**(6): 815–22.

Pelphrey, K. A., Mitchell, T. V., McKeown, M. J., Goldstein, J., Allison, T., & McCarthy, G. (2003). Brain activity evoked by the perception of human walking: Controlling for meaningful coherent motion. *Journal of Neuroscience*, **23**, 6819–25.

Pelphrey, K. A., Morris, J. P., Michelich, C. R., Allison, T., & McCarthy, G. (2005). Functional anatomy of biological motion perception in posterior temporal cortex: An fMRI study of eye, mouth and hand movements. *Cerebral Cortex*, **15**(12): 1866–76.

Pelphrey, K. A., & Morris, J. P. (2006). Brain mechanisms for interpreting the actions of others from biological-motion cues. *Current Directions in Psychological Science*, **15**(3): 136.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge: MIT Press.

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

Platek, S., Keenan, J. P., Gallup Jr, G. G., & Mohamed, F. B. (2004). Where am I? The neurological correlates of self and other. *Cognitive Brain Research* **19**(2): 114–22.

Poldrack, R. A. (2006a). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* **10**(2): 59–63.

Poldrack, R. A. (2006b). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience* **2**(1): 67–70.

Rabin, J. S., Gilboa, A., Stuss, D. T., Mar, R. A., & Rosenbaum, R. S. (2010). Common and unique neural correlates of autobiographical memory and theory of mind. *Journal of Cognitive Neuroscience* **22**(6): 1095–111.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14-and 18-month-olds. *Developmental Psychology* **33**(1): 12.

Samson, D., Apperly, I. A., & Humphreys, G. W. (2007). Error analyses reveal contrasting deficits in "theory of mind": Neuropsychological evidence from a 3-option false belief task. *Neuropsychologia* **45**(11): 2561–9.

Saxe R. (In press). The new puzzle of theory of mind development. In: M. R. Banaji & S. A. Gelman (Eds), *The Development of Navigating the Social Cognition. World: What infants, children, and other species can teach us*. New York: Oxford University Press.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind." *NeuroImage* **19**(4): 1835–42.

Saxe, R., Moran, J. M., Scholz, J., & Gabrieli, J. (2006a). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience* **1**(3): 229–34.

Saxe, R., & Offen, S. (2010). Seeing ourselves: What vision can teach us about metacognition. In: G. Dimaggio, & P. H. Lysaker (Eds), *Metacognition and Severe Adult Mental Disorders: From basic research to treatment* (pp. 13–29). New York: Taylor & Francis.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science* **17**(8): 692–9.

Saxe, R. R., Schulz, L. E., & Jiang, Y. V. (2006b). Reading minds vs. following rules: dissociating theory of mind and executive control in the brain. *Social Neuroscience*, **1**(3–4): 284–98.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia* **43**(10): 1391–9.

Saxe, R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Development*, **80**(4): 1197–209.

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**(3): e4869–e4869.

Schnell, K., Bluschke, S., Konradt, B., & Walter, H. (2011). Functional relations of empathy and mentalizing: An fMRI study on the neural basis of cognitive empathy. *NeuroImage* **54**(2): 1743–54.

Serences, J. T., Shomstein, S., Leber, A. B., Golay, X., Egeth, H. E., & Yantis, S. (2005). Coordination of voluntary and stimulus-driven attentional control in human cortex. *Psychological Science* **16:**114–22.

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. *Neuropsychologia* **45**(13): 3054–67.

Shamay-Tsoory, S., Tomer, R., Berger, B. D., Goldsher, D., & Aharon-Peretz, J. (2005). Impaired affective theory of mind is associated with right ventromedial prefrontal damage. *Cognitive and Behavioral Neurology* **18**(1): 55–67.

Slessor, G., Phillips, L. H., & Bull, R. (2007). Exploring the specificity of age-related differences in theory of mind tasks. *Psychology and Aging*, **22**(3): 639–43.

Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. *NeuroImage*, **35**(3): 1378–84.

Sommer, M., Rothmayr, C., Döhnel, K., Meinhardt, J., Schwerdtner, J., Sodian, B., & Hajak, G. (2010). How should I decide? The neural correlates of everyday moral reasoning. *Neuropsychologia*, **48**(7): 2018–26.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science* **18**(7): 587–92.

Spiers, H. J., & Maguire, E. A. (2006). Thoughts, behaviour, and brain dynamics during navigation in the real world. *NeuroImage*, **31**(4): 1826–40.

Spunt, R. P., & Lieberman, M. D. (2012). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. *NeuroImage* **59**(3), 3050.

Spunt, R., Satpute, A. B., & Lieberman, M. (2011). Identifying the what, why, and how of an observed action: an fMRI study of mentalizing and mechanizing during action observation. *Journal of Cognitive Neuroscience* **23**(1): 63–74.

Tamir, D. I. & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences*, **107**(24): 10827.

Tsakiris, M., Costantini, M., & Haggard, P. (2008). The role of the right temporo-parietal junction in maintaining a coherent sense of one's body. *Neuropsychologia* **46**(12): 3014–18.

Uttal, W. R. (2011). *Mind and Brain: A Critical Appraisal of Cognitive Neuroscience*. Cambridge: MIT Press.

Van Overwalle, F. (2008). Social cognition and the brain: A meta-analysis. *Human Brain Mapping* **30**(3): 829–58.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., … & Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage* **14**(1): 170–81.

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**, 817–27.

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain:: an event-related fMRI study. *Neuron*, **30**(3): 829–41.

Wagner, D. D., Kelley, W. M., & Heatherton, T. F. (2011). Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. *Cerebral Cortex*, **21**(12): 2788–96.

Walter, H., Schnell, K., Erk, S., Arnold, C., Kirsch, P., Esslinger, C., et al. (2010). Effects of a genome-wide supported psychosis risk variant on neural activation during a theory-of-mind task. *Molecular Psychiatry* **16**(4): 462–70.

Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3): 655–84.

Whitfield-Gabrieli, S., Moran, J. M., Nieto-Castañón, A., Triantafyllou, C., Saxe, R., & Gabrieli, J. D. (2011). Associations and dissociations between default and self-reference networks in the human brain. *NeuroImage*, **55**(1): 225–32.

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): 103–28.

Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three- dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism* **12**(6): 900–18.

Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *Journal of Neuroscience* **30**(32): 10744–51.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010a). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, **107**(15): 6753–8.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences* **104**(20): 8235–40.

Young, L., Dodell-Feder, D., & Saxe, R. (2010b). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, **48**(9): 2658–64.

Young, L., Nichols, S., & Saxe, R. (2010c). Investigating the neural and cognitive basis of moral luck: it's not what you do but what you know. *Review of Philosophy and Psychology*, **1**(3): 333–49.

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, **40**(4): 1912–20.

Young, L., & Saxe, R. (2009a). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience* **21**(7): 1396–405.

Young, L., & Saxe, R. (2009b). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia* **47**(10): 2065–72.

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and. *Cognition* **35**: 41–68.

Zaitchik, D., Koff, E., Brownell, H., Winner, E., & Albert, M. (2006). Inference of beliefs and emotions in patients with Alzheimer. *Neuropsychology* **20**, 11–20.

Zaitchik, D., Walker, C., Miller, S., LaViolette, P., Feczko, E., & Dickerson, B. C. (2010). Mental state attribution and the temporoparietal junction: An fMRI study comparing belief, emotion, and perception. *Neuropsychologia* **48**(9): 2528–36.