The right temporo-parietal junction: a specific brain region for thinking about thoughts

Rebecca Saxe

Department Brain and Cognitive Sciences, MIT

It is easy to miss an extraordinary moment while it is happening. In the early 2000s, access to non-invasive human neuroimaging was spreading, and researchers were gaining the confidence to take these tools beyond the range of traditional neuroscience. Until then, neuroimaging had mostly been used to visualize brain functions that were already known from neuropsychology or animal studies: motor cortex (Puce et al., 1995), early visual cortex (Engel, Glover & Wandell, 1997; Engel et al., 1994), the motion perception response in MT+ (Smith, Greenlee, Singh, Kraemer & Hennig, 1998; Tootell et al., 1995), left lateralized language regions (Binder et al., 1997; Binder et al., 1996; Pujol, Deus, Losilla & Capdevila, 1999), and so on. Around 2000, though, beginning in London (Castelli, Happe, Frith & Frith, 2000; Fletcher et al., 1995b; Gallagher et al., 2000) and then spreading around the world (Brunet, Sarfati, Hardy-Baylé & Decety, 2000; Saxe & Kanwisher, 2003; Vogeley et al., 2001), cognitive neuroscientists leapt into a new domain: Theory of Mind (ToM), or the ability to think about thoughts. At that point, something remarkable happened. Across different labs, countries, tasks, stimuli and scanners, every lab that asked, 'which brain regions are involved in ToM?', got the basically same answer: a group of brain regions in the right and left temporo-parietal junction (TPJ) , right anterior superior temporal sulcus (STS) and temporal pole, the medial precuneus and posterior cingulate (PC) and the medial prefrontal cortex (MPFC; (Castelli et al., 2000; Fletcher et al., 1995b; Gallagher et al., 2000; German, Niehaus, Roarty, Giesbrecht & Miller, 2004; Goel, Grafman, Sadato & Hallett, 1995; Saxe & Kanwisher, 2003; Vogeley et al., 2001).

Of course, there has never been so much consensus again. From the very first, the researchers disagreed about what these results *meant*. Which of these brains regions was specifically recruited for ToM, and which came along for the ride, perhaps because of confounds in the experimental design? How many different brain regions were there in the group? For example, was there a single region of activity in the medial prefrontal cortex, or three? What functional

contribution was each region making? In what other, perhaps non-social, cognitive tasks did these brain regions participate? The answers to all of these questions remain controversial.

The current chapter will describe evidence for one hypothesis: namely, that there is a region near the right temporo-parietal junction which plays a specific role in representing thoughts (e.g. people's beliefs, desires, and emotions, but not physical sensations), and not in any other cognitive task. I believe that this hypothesis is the best interpretation of evidence from dozens of studies conducted over the last decade, in my lab and in others. (The functions of the other brain regions in the group are also very interesting, but would require whole other chapters to discuss). Looking back, though, what remains most striking is the original consensus. Because there was almost no pre-existing neuroscience of ToM, cognitive neuroscience researchers came to the topic with unusually few preconceptions about where to look in the brain, or how (i.e. with what tasks). In those circumstances, neuroimaging is notoriously fickle, producing many false positives and false negatives. Yet every group that sought to identify brain regions implicated in ToM got essentially the same answer; and in study after study, we still do. Any interpretation must begin with that fact.

Why hypothesize a domain-specific neural mechanism for Theory of Mind?

The prediction that humans might have a distinct brain mechanism, specifically involved in ToM, originally arose from purely behavioral studies of children. The development of ToM is typically tested using a paradigm called the 'False belief task'. A child watches while a puppet places an object in location A. The puppet leaves the scene and the object is transferred to location B. The puppet returns and the child is asked to predict where the puppet will look for the object. Performance on this task changes dramatically in children, between ages three and five years. Three-year-olds predict the puppet will look in location B, where the object actually is; five-year-olds predict the puppet will look in location A, where the puppet last saw the object (Wellman, Cross & Watson, 2001). Notable, the three-year-olds are not performing at chance, or confused by the questions; they make systematically below-chance predictions, with high confidence (Ruffman, Garnham, Import & Connolly, 2001). This same pattern of change has been observed in hundreds of studies, and across many cultures (Lee, Olson & Torrance, 1999;

Liu, Wellman, Tardif & Sabbagh, 2008; Wellman et al., 2001; Yazdi, German, Defeyter & Siegal, 2006), from wealthy industrialized societies to hunter-gatherers (Avis & Harris, 1991). Thus, the early stages of ToM apparently follow a universal developmental trajectory.

There is a revealing exception to this pattern, though. Children diagnosed with autism spectrum disorders (ASD) are significantly delayed in passing false belief task, compared to typically developing children or to children with other developmental disorders like Down Syndrome (Baron-Cohen, 1989; Baron-Cohen, Leslie & Frith, 1985). On a larger set of tasks, tapping multiple different aspects of ToM (e.g. understanding desires, beliefs, knowledge and emotions), children with ASD show both delayed development and also disorganized development. That is, typically developing children pass these tasks in a stable order (e.g. understanding false beliefs is easier than understanding false emotions), but children with autism pass the tasks in a scrambled order, as if they were passing for different reasons (Peterson, Wellman & Liu, 2005).

The observation that a neurobiological developmental disorder, ASD, could disproportionately affect development of ToM, led researchers to hypothesize that ToM development depends on a distinct neural mechanism. That is, some brain region, chemical, or pattern of connectivity might be specifically necessary for ToM, and disproportionately targeted by the mechanism of ASD. This hypothesis was difficult to test, though, until the advent of neuroimaging allowed researchers to investigate the brain regions underlying high-level cognitive functions like ToM.

Discovering the RTPJ: a response specific to ToM?

Following the tradition in developmental psychology, most of the early neuroimaging studies of ToM required participants to attribute false beliefs to characters in stories or cartoons. Meanwhile, the scientists measured the flow blood carrying oxygen around the participant's brain, either by putting radio-active labels in the blood (positron emission tomography, PET; Happé et al., 1996) or by measuring intrinsic differences in the magnetic response of oxygenated blood (functional magnetic resonance imaging, fMRI; Brunet, Sarfati, Hardy-Bayle & Decety, 2000; Fletcher et al., 1995a; Gallagher et al., 2000; Goel et al., 1995; Saxe & Kanwisher, 2003; Vogeley et al., 2001). As I mentioned above, a reliable set of brain regions showed increased

oxygenation (implying increased metabolism, or "activity") during the false belief condition of each study, including a region in the right TPJ. The controversy that arose next was not over *whether* a region in the right TPJ was recruited during false belief tasks, but *why*.

I propose that there is activity in the right TPJ during false belief tasks because this region plays a specific role in thinking about thoughts. On this view, the right TPJ contains domain-specific representations with content like "Bob thinks that his keys are in his pocket" and "Alice wants to get a PhD from MIT." These representations form the core of ToM. Of course, the RTPJ cannot form or use such representations on its own, without input from and output to other brain regions. My suggestion is that the right TPJ contains the domain-specific content of these thoughts - the part that lets us explicitly conceive of "thinking" or "wanting" something to be true - and then interacts with many other brain regions in order to use these ideas in context.

Before this hypothesis can begin to sound plausible, there are a lot of other alternative hypotheses that we must rule out. There is more to solving a false belief task than a concept of "belief" (or "mental state" or "propositional attitude"), and there is more to a concept of belief than passing the false belief task. So in this section, I will show that it is specifically the need to think about thoughts, and not any other aspect of solving a false belief task, that is the best predictor of activation in the right TPJ.

Alternative 1: Inhibitory control

The first alternative hypothesis is that activity in the right TPJ reflects the structural demands of the false belief task, not the content. In the standard false belief task, described above, the participant has to juggle two competing representations of reality (the actual state of affairs versus the world as represented in the character's head) and then inhibit an incorrect but compelling answer (the true location of the object), in order to respond based on the character's beliefs. Performance on the false belief task thus depends as much on children's inhibitory control and executive function as it does on their ability to think about thoughts.

Three independent lines of research suggest that response selection and inhibitory control play a non-trivial role in the false belief task. First, 3-year-olds are more likely to pass versions of the task that reduce the demands on inhibition (e.g. Lewis & Osborne, 1990; Mitchell & Lacohée,

1991; Wellman & Bartsch, 1998; Wellman, Cross & Watson, 2001; Yazdi, German, Defeyter & Siegal, 2006). Second, individual differences in executive control are strongly correlated with individual differences in children's performance on theory of mind tasks (Carlson & Moses, 2001; Sabbagh, Fen, Carlson, Moses & Lee, 2006). Third, young children who fail the false belief task also fail control tasks that place comparable demands on executive function but do not include any reference to other minds or mental states (Roth & Leslie, 1998; Slaughter, 1998; Zatichik, 1990). For example, reasoning about false or misleading signs and maps is similar to reasoning about false beliefs: in both cases, the representation (belief or sign/map) is designed to correspond to reality, but fails to do so because of a mistake or because it is outdated (e.g. a sign indicates that a pie contains apple, and Mary believes that the pie contains apple, whereas in reality the pie contains pears). If asked "what does Mary think is in the pie?" or "what does the sign say is in the pie?", participants must consider two competing responses (e.g. apples versus pears), and inhibit the reality (pears), in order to produce the response based on the false representation (apples). In the false sign/map task, no mental state understanding is required but 3-year-olds fail the task.

Executive function is not sufficient for false belief task performance, though. Age-related deficits on the false belief task persist even when executive demands are substantially reduced (Wellman et al., 2001). Children under three fail even the simplest versions of the task (though see recent work by (Onishi & Baillargeon, 2005). Also, group differences in executive control and Theory of Mind skills can be dissociated. For example, although performance on executive tasks is correlated in individuals in a single culture, Chinese preschoolers out-perform American children on every test of executive control, but show no advantage on standard Theory of Mind tasks (Sabbagh et al., 2006). Executive control and Theory of Mind skills may also be dissociable in neuropsychological populations, such as children with autism. Children with autism do who fail false belief tasks nevertheless perform well on the false sign/map task (Charman & Baron-Cohen, 1992; Leslie & Thaiss, 1992).

So development evidence suggests that both executive function and ToM are necessary, and neither is sufficient, for success on the standard false belief task. A key question is therefore: which of these kinds of cognitive demands predicts activity in the right TPJ? Could it be that the

right TPJ is recruited for selecting among competing alternative representations, regardless of whether those representations include anyone's thoughts or beliefs?

These alternatives can be tested directly, by measuring right TPJ activity while participants reason about false physical representations, like signs and maps. The difference between false beliefs and false signs is just that in the false belief task, participants have to think about *mental* representations - i.e. thoughts. For the right TPJ, this difference makes the difference: activity in the right TPJ is high while participants are thinking about false beliefs, but no different from resting levels while participants are thinking about false photographs, maps or signs (Perner, Aichhorn, Kronbichler, Staffen & Ladurner, 2006; Saxe & Kanwisher, 2003). That is, if the participant is reading about photographs or maps, if you looked just at activity in the right TPJ, you wouldn't know there was anything on the participant's screen at all.

This pattern isn't true everywhere in the brain. Dorsolateral prefrontal and superior parietal brain regions show equally high activity in both the false belief and false photograph tasks (Saxe et al., 2006). These brain regions seem to play a general role in choosing among competing alternatives: the same regions also show high activity in many other tasks that demand response selection (e.g. Bunge, Hazeltine, Scanlon, Rosen & Gabrieli, 2002; Corbetta, Shulman, Miezin & Petersen, 1995; MacDonald, Cohen, Stenger & Carter, 2000; Rowe, Toni, Josephs, Frackowiak & Passingham, 2000). The neuroimaging results converge with neuropsychological evidence showing that patients with dorsolateral prefrontal lesions have general problems with executive function - especially when they need to choose between alternative answers or responses (e.g. Samson, Apperly, Kathirgamanathan & Humphreys, 2005; Stuss & Benson, 1984). Dorsolateral prefrontal regions thus appear to contribute to the domain-general demands of the false belief task, i.e. the cognitive demands that the false belief task shares with the false photograph task.

Interestingly, another pattern emerges in the left TPJ. The left TPJ response is higher for both false beliefs and false signs than for any other control condition (Perner et al., 2006). Again, this fits with neuropsychological evidence: patients with damage to the left TPJ fail both false belief and false photograph tasks, but not other tasks that tap inhibitory control or response selection (Apperly, Samson, Chiavarino, Bickerton & Humphreys, 2007). The left TPJ can be interpreted as playing a role in "meta-representation": helping people think about the idea of a representation

(Perner, 1991). Again, this function is not specific to ToM (Apperly, Samson & Humphreys, 2005).

The right TPJ, though, does seem to be specific to ToM. Activity in the right TPJ is predicted by the need to think about thoughts, not by the need to select among competing responses or form a meta-representation. Passing a false belief task doesn't depend only on the right TPJ; but it might depend especially on the right TPJ. Still, we need to rule out a second alternative hypothesis.

Alternative 2: Any information about people

Humans live in a social world. Correspondingly, there are many different human brain regions that are involved in perceiving and thinking about other people: brain regions for recognizing people's faces (Kanwisher, McDermott & Chun, 1997), facial expressions (Haxby, Hoffman & Gobbini, 2002; Hoffman & Haxby, 2000), voices (Belin, Zatorre, Lafaille, Ahad & Pike, 2000), bodies (Downing, Jiang, Shuman & Kanwisher, 2001) and actions (Decety & Grezes, 1999; Decety et al., 1997). The second alternative hypothesis is therefore that the right TPJ contains an abstract representation of other people - or perhaps of socially relevant facts about people - but not specifically about other people's thoughts. This hypothesis seemed especially plausible when the right TPJ was first discovered because of its anatomical position (e.g. Castelli, Frith, Happe & Frith, 2002; Frith & Frith, 1999; Frith & Frith, 2001). The right TPJ is located posterior to the ascending branch of the right superior temporal sulcus, very close to regions of the STS that show activity for all kinds of social stimuli, including images of people's facial expressions and bodily movements. (For more on the right STS, see below). The right TPJ activity might therefore be a part of this broader region's response to people, not specific to people's thoughts. (When I first started working on the right TPJ, this was actually my preferred hypothesis; I predicted that the right TPJ was a person- or agent-detection mechanism. Like almost everyone else, I expected that true ToM would depend on regions in the frontal lobe).

To test this hypothesis, we have measured the response of the right TPJ to a range of stimuli containing people, but not information about the people's thoughts. The right TPJ does not respond to photographs of people (Saxe & Kanwisher, 2003) or to descriptions of people's physical appearance (Saxe & Kanwisher, 2003; Saxe & Powell, 2006). Activity in the right TPJ

is also low during descriptions of people's physical sensations like hunger, thirst, or tiredness (Bedny, Pascual-Leone & Saxe, 2009; Saxe & Powell, 2006).

Even within a single story, the timing of the response in the RTPJ is predicted by the timing of sentences describing a character's thoughts. Blood oxygen levels change slowly; it takes 4 -6 seconds for the blood oxygen level to reach its peak, even when the underlying neurons' electrical activity lasts only a hundred milliseconds. As a consequence, blood oxygen based neuroimaging tools, like PET and fMRI, have poor temporal resolution. Still, it's possible to distinguish the neural response to events that happen at least six seconds apart. In three studies, we have done just that: information about a character's thoughts is presented in 6-20 second chunks within a single ongoing story. Across all three studies, the response in the right TPJ shows a peak just at the time when someone's thoughts are described (Saxe, Whitfield-Gabrieli, Scholz & Pelphrey, 2009; Saxe & Wexler, 2005; Young, Cushman, Hauser & Saxe, 2007).

So the right TPJ contradicts our second alternative hypothesis: general social information about people gets no response. Activity in the right TPJ is predicted by whether, and when, a stimulus describes a person's thoughts - at least in adults. The response pattern in young children is different; more on this below.

The RTPJ and its neighbors

The hypothesis of this chapter is that a region in the right TPJ plays a specific role in thinking about thoughts. So far, I've mentioned eight different experiments whose results fit this hypothesis. All of those experiments use one kind of statistical analysis, called a functional region of interest (or fROI) analysis. That is, the researchers first identify the region in the right TPJ that shows a high response during the false belief task in each individual participant's brain, and then ask how *that* region responds in some new conditions. My lab almost always uses fROI analyses, and the results are consistent: the right TPJ response is specific to thinking about thoughts. Many other cognitive neuroscientists, however, use a different kind of statistical analysis, called whole-brain group analyses, and seem to get a different result: right TPJ activation shows up in other tasks that seem to have nothing to do with ToM. What is going on?

The non-social task that most frequently elicits activity in right TPJ, in whole-brain group analyses, is called 'exogenously cued attention': basically, shifting attention to an unexpected stimulus in the environment. Exogenous attention is contrasted with endogenous attention, when the person's own goals determine which stimuli in the environment will get his or her attention. For a sense of the distinction, imagine you're walking through Time Square looking for a blind-date who said he would be wearing a green shirt: your attention will be directed towards green things, endogenously, but also, whether you like it or not, to the bright flashing billboards, exogenously. Many fMRI studies find that when participants shift their attention to unexpected stimuli - whether the stimuli are unexpected in time (e.g. low-frequency targets) or space (e.g. invalidly cued locations) - there is increased activity in a region of the right TPJ (Bledowski, Prvulovic, Goebel, Zanella, & Linden, 2004; Downar, Crawley, Mikulis, & Davis, 2000; Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000; Vossel, Weidner, Thiel, & Fink, 2009). These results are also consistent with neuropsychological evidence. People with strokes that damage the right TPJ show an impairment in reorienting attention (Friedrich, Egly, Rafal, & Beck, 1998), often resulting in left-hemisphere neglect (Vallar, Bottini, Rusconi, & Sterzi, 1993; Vallar & Perani, 1986).

Why would the right TPJ be activated for both thinking about thoughts, and shifting attention? The answer to this question depends on the answer to a prior question: do these two tasks lead to activity in the *same* region in the right TPJ? There is a lot of cortical space near the intersection of the temporal and parietal lobes - enough for tens of millions of neurons. Perhaps we have simply been confusing two neighboring but distinct groups of neurons, one involved in ToM and another involved in exogenous attention. This is why I mentioned the two different analysis strategies that are common in cognitive neuroscience: whole-brain group analyses are much more likely than fROI analyses to blur the boundaries between neighboring brain regions, creating the appearance of overlap where there are really two distinct functions.

FROI and whole-brain group analyses are different solutions to the same problem: how to identify the *same place* in different people's brains. Individuals' brains are as different as their faces: like faces, brains have major anatomical landmarks in common (eyes, nose; STS, occipital pole), but differ in the shapes, sizes, and relative positions of the parts and of the whole. Fortunately for neuroscience, the large-scale functional structure of the brain is also shared

across people. Basic visual functions, for example, occur along the calcarine sulcus at the occipital pole; motor commands come from the motor homunculus, anterior to the central sulcus. But these sulci are slightly different shapes, sizes, and distances from each other in each individual; and the size and shapes of the functional regions also differ. So the challenge is how to allign these brains to one another, so that we can study the function of some particular region in each brain.

Group analyses stretch and warp each brain, to try to allign it to a single common shape - an average brain shape. If this works, then everyone's individual visual cortex will end up in the position of the average visual cortex, and everyone's individual motor cortex will end up in the position of the average motor cortex, and so on. Many different algorithms have been developed for this process - called 'normalising' the brain. Some try to allign every individual crease and fold (sulcus and gyrus) in the cortex (e.g. by normalising the 'surface' of the cortex); the more common algorithms try to match just the overall shape of the brain (the 'volume'). All of the modern algorithms do a decent job, and some of them can do an amazing job of alligning at least some regions (Amunts, Malikovic, Mohlberg, Schormann & Zilles, 2000; Hinds et al., 2008). But where the algorithm is imperfect, then region X in one person's brain will end up in the same position, in the average brain, as the neighbouring region Y in another person's brain. As a result, the group average region will appear to have the functions of both X and Y, even if X and Y were anatomically distinct in every individual. (This problem is exacerbated in whole-brain meta-analyses, in which the position of X was measured in one group of brains, and the position of Y was measured in a different group of brains).

As I mentioned, fROI analyses take a different approach: they use a 'marker' function to identify the region of interest in each individual brain, and then study it. As long as the 'marker' function reliably picks out the same functional region across people, this technique can avoid all the problems of functional and anatomical heterogeneity. As a consequence, fROI analyses are always used by researchers studying brain regions with well-understood marker functions - like retinotopy in early visual cortex.

I've taken this long detour into fMRI analysis techniques, because I think considerable confusion about the function of the right TPJ can be explained by the bias in whole-brain group analyses to

blur together neighbouring regions. When we measure the positions of the regions involved in ToM, and in exogenous attention, in the same individuals, we find two neighbouring but distinct regions (Scholz, Triantafyllou, Whitfield-Gabrieli, Brown & Saxe, 2009). The activity elicited by exogenous attention is reliably 10 mm superior to the activity elicited by ToM. From one brain to another, the position of each of these brain regions moves around by more than 10 mm, but their relative positions in one individual brain is reliable: exactly the kind of situation in which two distinct brain regions appear to overlap in group analyses.

A similar confusion existed for a while between the TPJ and its anterior neighbour, the right posterior STS. Initially, many researchers concluded that there was one brain region in this vicinity, sometimes called the TPJ/STS, involved in understanding other people's actions. More recent research has clarified that these are different brain regions: the right TPJ is recruited during stories or cartoons that depict a person's thoughts, but not for other depictions of human actions, whereas the right posterior STS is recruited when watching movies of simple actions, like reaching for a cup, but not for verbal descriptions of actions. These regions are anatomically distinct: the right TPJ is posterior to the STS in both adults (Gobbini, Koralek, Bryan, Montgomery & Haxby, 2007) and children (Saxe et al., 2009). Another interesting functional distinction emerges with repetition: the STS continues to respond even if the same individual movie is played eight or sixteen times (Brass, Schmitt, Spengler & Gergely, 2007; Hamilton & Grafton, 2006; Pelphrey, Morris & McCarthy, 2004), while the TPJ response is massively reduced on just the second presentation of the same story (Andrews-Hanna, Chakroff, Bucker, & Saxe, in prep). That is, the STS response seems to be bound to the stimulus, so the same stimulus elicits the same response, over and over again; by contrast, the TPJ response seems to reflect something more like an *inference* from the stimulus, which occurs only on the first encounter.

The existence of at least three different functional regions in a small(-ish) area of cortex is not surprising, but it does create confusion - especially in nomenclature. How should we distinguish between the anatomical area at the junction between the temporal and parietal lobes, and the functional regions in those anatomical areas, involved in action perception, exogenous attention, and ToM, respectively? In current literature, for maximum confusion, the term "right STS/TPJ" is frequently used for all of them. In other parts of the brain, this problem was (temporarily) avoided by giving functional regions names that distinguished them from their broader

anatomical location (e.g. the Fusiform Face Area, or FFA, is a functionally-defined subset of the fusiform gyrus, (Kanwisher et al., 1997). Even this solution was only temporary, though, because higher-resolution scanning is now revealing multiple different sub-regions within the FFA, whose functions and names have yet to be determined.) I tried to take a similar strategy by dubbing the functionally-defined subset of the right TPJ that is involved in ToM the right TPJ-M (for "Mental"; Saxe, 2006; Saxe & Kanwisher, 2003). But the name never caught on.

A third alternative: thoughts about thoughts, or words about thoughts?

Distinguishing the right TPJ region involved in ToM (hereafter, just "right TPJ" again) from its neighbours helps clarify the literature, and revives the possibility that this brain region is involved selectively in thinking about thoughts. We are not yet finished testing alternative hypotheses, though. In all of the experiments I have described so far, participants are induced to think about thoughts when they read, or hear, a sentence that explicitly describes a thought (e.g. "Bob knows his flight is delayed" or "Grace thinks the white powder is sugar"). So it's possible (and this is the third alternative hypothesis) that the right TPJ is involved not in *thinking* about thoughts, but in understanding words or sentences that describe thoughts.

These questions about the right TPJ are connected to a broader theoretical question about the relationship between language and ToM: namely, is it possible to think about thoughts without thinking in sentences? Maybe ToM depends entirely on a special kind of linguistic structure. Sentences about thoughts have unusual structural properties: first, they contain another sentence embedded inside themselves (e.g. "the chocolate is in the top box"), and second, the truth of the whole sentence does not depend on the truth of the embedded sentence (because the girl's belief may be false). Perhaps people can formulate thoughts about thoughts, including false beliefs, only by relying on our linguistic ability to construct this kind of sentence. There is some reason to think so.

During development, general linguistic ability strongly predicts ToM task performance, even on non-verbal tasks (e.g. Astington & Jenkins, 1999; Dunn & Brophy, 2005). Children's production and comprehension of the specific structure of ToM sentences precedes and predicts performance on the false belief task, in both correlational and training studies (Hale & Tager-

Flusberg, 2003; de Villiers, 2000; de Villiers & de Villiers, 2000) but see (Bartsch & Wellman, 1995; Ruffman, Slade, Rowlandson, Rumsey & Garnham, 2003). Also, language development seems to be necessary and sufficient for ToM development. Deaf children of non-signing parents get reduced exposure to language in the first year or two of life, and are selectively delayed in passing ToM tasks (e.g. Peterson & Siegal, 1999; Wellman & Liu, 2004). This delay is caused by delayed language exposure, not by being deaf; deaf children of native signers - who learn sign language from birth - show no delay at all (de Villiers, 2005). Adults who speak an emerging sign language without any mental state terms perform poorly on false belief tasks; but when those same adults learn a new mental state vocabulary, their performance on the false belief task improves (Pyers & Senghas, 2009). Even adult native English speakers perform poorly on a false belief task if their ability to use language during the task is temporarily blocked ((Newton & de Villiers, 2007).

Other evidence, though, suggests it *is* possible to think about thoughts without using language. Adults who acquired ToM typically, during development, and then suffer catastrophic loss of language capacities due to left hemisphere stroke, do not lose ToM (Siegal & Varley, 2006). For example, patient PH is severely impaired on all aspects of syntax, following a massive left hemishere stroke (Apperly, Samson, Carroll, Hussain & Humphreys, 2006). He cannot even use grammar to understand 'Mary was pushed by Bob'; he fails completely on embedded sentences. Nevertheless, PH performs perfectly well on non-verbal tests of ToM, including complex 2nd-order inferences (what X thinks that Y thinks). So with practice, it seems that people can think about thoughts using non-linguistic strategies.

One way to test this hypothesis is to ask whether activity in the right TPJ is predicted by the presence of mental state words in the stimulus. The simple answer is: no, it isn't. On the one hand, it is possible to elicit robust activity in the right TPJ without any words in the stimuli. The right TPJ responds robustly to non-verbal single-frame cartoons that depict someone having a false belief (Gallagher et al., 2000). The right TPJ also responds robustly to verbal descriptions of actions that imply a false belief without explicitly mentioning any thoughts (Saxe & Kanwisher, 2003, Experiment 1). On the other hand, "mental words" on their own do not elicit activity in the right TPJ. Neither mental adjectives (like 'curious', 'suspicious'; Mitchell, Banaji

& Macrae, 2005), nor mental state verbs (like 'to think', 'to suspect'; Bedny, M. personal communication) provoke right TPJ activity, when presented alone.

Even more compelling, differential activity in the right TPJ depends on whether a person is thinking about thoughts, even for the very same stimulus. For example, in one experiment the stimulus is an ambiguous stick-figure animation is disambiguated by the task instructions. One task requires participants to treat the stick-figure as a physical cue (basically like an arrow, the 'algorithm' task), while the other task asks participants to treat the stick-figure as a person, and to consider that person's thoughts (the 'ToM' task). Although both the stimuli and the responses are identical across tasks, right TPJ activity is higher when participants are doing the ToM task (Saxe et al., 2006).

From these results, it's clear that activity in the right TPJ is not predicted by mental state language in the stimulus. Still, it's hard to completely rule out our third alternative hypothesis. Right TPJ activity might be predicted by mental state language in the participant's head. That is, when asked to think about thoughts, people might do so by mentally formulating corresponding sentences. So when looking at a non-verbal cartoon, or an ambiguous stick-figure, people might be thinking in words, in their native language, something like "Well, the girl thinks the chocolate is in the top box."

So it remains an open question whether activity in the right TPJ reflects verbal or non-verbal representations of thoughts. Fortunately, this question can be tested empirically. One way would be to scan a patient with a left-hemisphere stroke, like PH, while he is solving non-verbal false belief tasks. I predict that when PH is using ToM, he uses his right TPJ just like the rest of us. If so, that would be the strongest evidence so far that activity in the right TPJ specifically reflects thinking about thoughts.

Moral judgment and 'reverse' inferences.

Once we establish that activity in the right TPJ reflects specifically thinking about thoughts, then we will be able to turn the tables: instead of assuming a task requires ToM and testing whether it produces activity in the right TPJ, we will be able to use activity in the right TPJ as a barometer

of ToM and test whether and when people are thinking about thoughts. Using brain activity to infer cognitive activity is called a 'reverse inference'. Initially, reverse inferences may seem circular. In order to establish that the right TPJ is responsible for ToM, we must establish that it is recruited for all and only the tasks that involve ToM; but sometimes, the best evidence that a task does invoke ToM may be the level of activation in the right TPJ. Fortunately, this circularity is temporary. Different theories of ToM and different theories of the function of the right TPJ can be tested simultaneously. The correct resolution will be determined by the consistency of the data that emerge from these tests, and the richness of the theoretical progress those data support.

Along these lines, we have been simultaneously investigating the role of the right TPJ, and the role of ToM, in making moral judgments. When we morally evaluate an action, we typically consider not just what happened (the consequences) but what the person was thinking. For example, we blame people who attempt but fail to harm others (negative intentions, neutral consequences), while generally forgiving people who harm others accidentally and unknowingly (neutral intentions, negative consequences; Cushman, Young & Hauser, 2006; Young et al., 2007). fMRI evidence is beginning to paint a rich and complicated picture of the neural processes that support these judgements.

Activity in the right TPJ appears to reflect at least two different ways that thinking about thoughts is invoked during moral judgment. First, while reading a story that contains a description of a character's belief, participants form an initial representation of that belief. We call this process 'encoding' because it occurs when the belief is initially presented. The right TPJ shows very high activity during encoding, with a peak just when the participant is reading about a belief; but this response does not distinguish between different kinds of belief (e.g. morally negative versus neutral) and does not predict subsequent moral judgments (Young & Saxe, 2008). Later, though, the right TPJ shows a different pattern of response. After the whole story is presented, once participants begin making moral judgments, there is a second peak in right TPJ activity. We call this second phase 'integration', since it appears to reflect participants' *use* of previously encoded beliefs in calculating an overall moral judgment. At this phase, activity in the rTPJ does predict participants judgments. For example, people differ in their moral judgments about accidents: some pay more attention to the character's innocent beliefs and intentions, and so make relatively lenient moral judgments, while others pay more attention to the harmful

outcome of the action, and so make relatively harsh moral judgments. These individual differences in moral judgments are correlated with differences in the right TPJ during 'integration': there is more right TPJ activity in people who make more lenient judgments, and less right TPJ activity in people who make harsher moral judgments, in response to the same stories (Young & Saxe, 2009b).

So far so good for our reverse inferences. Peaks in right TPJ activity can be interpreted as peaks in thinking about thoughts; these peaks can be further sub-divided into different hypothesised cognitive processes, like 'encoding' and 'integration'; and activity during one of these processes can be used to predict people's moral judgments.

Other patterns in the right TPJ response, though, defy such neat interpretation. Specifically, the right TPJ response during 'integration', averaged across people, depends on *when* belief information was presented earlier in the story. If the character's beliefs are explicitly stated near the end of the trial, the right TPJ response at integration is highest when participants are reading about a failed attempt (negative beliefs, neutral consequences). On the other hand, if the character's beliefs are explicitly stated near the beginning of the trial, the right TPJ response at integration is highest for any story involving a neutral belief (accidental harms, and neutral actions). We have replicated each of these patterns in multiple different experiments, with different groups of participants (Young & Saxe, 2008; Young & Saxe, 2009a). If we follow our strategy of reverse inference, these results suggest that the cognitive process of thinking about thoughts, during integration, somehow depends on the order of information presented earlier in the trial. Precisely how or why this happens, though, is still not clear. Hopefully, future progress in either the neuroscience of the right TPJ, or the cognitive theory of ToM and moral judgment, or both, will provide a clarification.

Inside the right TPJ

The idea that a pattern of firing in a group of neurons can represent something as complex as an attributed thought is pretty astonishing. How are these neurons doing it? All of the experiments I've described show that activity in the right TPJ is high when the person is thinking about

thoughts, and low otherwise. The next phase of this research program must be to break "thinking about thoughts" into its component parts. Within visual brain regions, sub-groups of neurons respond to features or dimensions of the visual image: line orientations, spatial frequencies, colours, retinal positions. What are the features or dimensions of a represented thought? Unfortunately, for the most part, we don't know yet.

We have tried measuring the right TPJ response to descriptions of beliefs with different features: true versus false beliefs, positively-valenced versus negatively-valenced beliefs, beliefs with good consequences versus bad consequences, beliefs based on good reasons versus bad reasons, beliefs that caused actions versus emotions, and beliefs based on visual versus auditory experiences (Bedny et al., 2009; Powell, unpublished; Young & Saxe, 2008; Young et al., 2007). Most recently, we have compared activity in the right TPJ while people read about common-sense beliefs ("John believes that swimming is a good way to cool off") or about absurd beliefs ("John believes that swimming is a good way to grow fins"; Young, Dodell-Feder & Saxe, in press). None of these features make any difference; the average right TPJ response is equally high for all of them.

For true versus false beliefs, though, there appears to be a contradiction in the literature. In a non-verbal action prediction task, right TPJ activity is higher on false belief than true belief trials (Sommer et al., 2007), but for verbal stimuli, right TPJ activity is equally high for true and false beliefs (Young et al., 2007). What explains these conflicting results? My interpretation is that the differential response reflects not the kind of belief, but whether or not participants were thinking about beliefs at all. When the belief is stated explicitly in the story, participants have no choice about whether to think about the belief, whether it's true or false. Non-verbal action prediction tasks are different. False belief trials require the participants to use the character's false beliefs to predict the character's actions (e.g., looking for an object in the wrong place). By contrast, true belief trials do not require belief attributions at all; participants may simply respond based on the true location of the object.

So the current literature seems consistent to me: average right TPJ activity does not differentiate between different kinds of attributed thoughts. To discover the dimensions or components of the neural contribution to thoughts about thoughts, we will have to use a different dependent

measure. In fact, it may be necessary to abandon fMRI, and switch to tools with higher spatial and/or temporal resolution like electroencephalograms, magnetoencephalograms, or direct electrical recording from neurons. I have not given up hope in fMRI yet, though. Higher resolution images (Grill-Spector, Sayres & Ress, 2006), and creative analysis techniques like multi-voxel pattern analysis (Haxby et al., 2001) or functional adaptation (Grill-Spector & Malach, 2001), may yet reveal the finer-grained structure of the representations inside the right TPJ.

Developing a right TPJ

Let's say my hypothesis is correct, and the right TPJ plays a specific role in thinking about thoughts. We would immediately confront the next big challenge: how does such a brain region develop? What combination of biological and environment factors lead a group of neurons to take on such a specific and abstract cognitive function? In my opinion, understanding the development of this brain region, and how it relates to the development of ToM, is the next major frontier for the cognitive neuroscience of ToM.

One way to investigate the factors driving typical development of the right TPJ is to measure the right TPJ in people who have had atypical experiences during development. For example, congenital blindness might affect the development of the right TPJ for any of three reasons: (i) because typical children learn about other minds first from visually observing human actions, (ii) if the neural representation of mental states is located in the TPJ due to connections with the STS representations of actions, and these connections might be disrupted by blindness, and/or (iii) specifically for reasoning about mental states based on an experience of sight, because congenitally blind adults have no first-person experience of sight themselves, which might influence their ability to attribute sight to others. In spite of all these possible sources of difference, the right TPJ in congenitally blind adults is identical in every respect we tested to the right TPJ of sighted adults (Bedny et al., 2009). That is, development of the right TPJ is largely resilient in the face of vast changes in both the modality of experience of other people, and in one's own first-person experiences. Thinking about other people's thoughts, in both sighted and

blind people, seems to depend on abstract representations that can be acquired without direct first-person experience.

Another approach is to investigate the development of the right TPJ during childhood, while ToM is first developing. Most research on the development of ToM has focused on early childhood: children aged 1 to 5 years. Over these years, children slowly master a whole foundation of mental state concepts, including beliefs, desires, knowledge, ignorance, and basic emotions. If ToM development has anything to do with maturation of the right TPJ, these must be key years for the right TPJ as well. Unfortunately, children so young are very hard to study with fMRI. An MRI scanner is a big noisy unfamiliar machine; and fMRI experiments require that participants stay perfectly still for at least 15 - 20 minutes. To date, the only way researchers have been able to keep very young children still for that long is to study them while they are asleep (Dehaene-Lambertz, Dehaene & Hertz-Pannier, 2002; Redcay, Haist & Courchesne, 2008; Redcay, Kennedy & Courchesne, 2007); but it is hard to imagine how to get sleeping children to engaged in ToM. So for now, studying the development of the right TPJ in children has required other strategies.

One approach is to switch to different neuroimaging methods, which are more child-friendly. Electroencephalograms (EEG) offer one such alternative. EEGs measure electrical fields on the scalp, the aggregate of all the tiny electrical fields produced by neurons when they fire. The spatial distribution of EEG signals is very blurry: activity from many different neurons and regions cancel each other out, or average together. Still, when a group of neurons are aligned, and firing synchronously, the resulting average electrical activity can be detected. EEGs can be analysed either (1) by looking at the evolution of the eletrical activity over time (e.g. ~500 milliseconds) after a pre-determined event (called 'event-related potentials', ERPs, (Luck, 2005)), or (2) by looking at the power and coherence of the EEG signal in its component frequencies ('spectral analyses'). ERPs measure exactly when activity occured after a stimulus; while spectral analyses measure groups of neurons that are firing synchronously in a periodic rhythm over time. Both kinds of analyses reveal developmental changes in brain organisation over the first few years of life (Courchesne, 1977; Courchesne, 1978; Courchesne, Ganz & Norcia, 1981; de Haan, Johnson & Halit, 2003; Holcomb, Coffey & Neville, 1992; Hyde, Jones, Porter & Flom, 2010; Sabbagh & Taylor, 2000).

Spectral analyses of EEG data suggest a link between the right TPJ and ToM development. Four year old children are just on the threshold of passing standard false belief tasks: some 4-year-olds consistently make action predictions based on the character's false belief, while others still consistently make predictions based on the reality. What makes the difference? Two regions show higher neural coherence at rest (in the 'alpha' frequency band, with a periodic cycle of 6-9 Hz), in the children who pass false belief tasks: the right TPJ and medial prefrontal cortex, two of the key regions for ToM in adults! So these data suggest that anatomical maturation of the right TPJ is connected with cognitive development of ToM.

On the other hand, ERP analyses - in both children and adults - point to a completely different neural basis of ToM: a late slow response in the left lateral frontal lobe appears while people are answering questions about false beliefs (Liu, Sabbagh, Gehring & Wellman, 2004; Liu, Sabbagh, Gehring & Wellman, 2009). What explains these conflicting results? I think the explanation lies in the way the false belief task is adapted for ERPs. ERP analyses require that the cognitive process of interest has a precise (to the millisecond) and predictable starting time. When does thinking about thoughts start? In a typical false belief task, people might begin to consider the character's mental states when the character turns his back, when the object is moved from its original location to a new location, when the character returns to the room, or when participants hear the question and have to formulate a response - or, as we saw in the case of moral judgment, some combination of all of the above. ERP studies of ToM, so far, focus on the cognitive processes right after participants see the question. So one possibility is that this time-window is just too late to see right TPJ activity: people may preemptively form (or 'encode') representations of the character's beliefs as the trial unfolds. At the time of the question, other cognitive and neural processes are required for solving false belief tasks, including (as described above) a left-lateral frontal region involved in inhibitory control - which may be the source of the ERP signal that's been reported in this literature. If this interpretation is right, future ERP studies that focus on neural activity earlier during the trial should reveal right TPJ activity in adults - and possibly the development of the right TPJ in children.

In the meantime, another approach to studying ToM development is to use fMRI in slightly older children. Although most studies of ToM development have focussed on children 5 years old and younger, ToM development does not actually stop at 5 years. More subtle aspects of ToM

continue to develop throughout childhood - including the use of beliefs and intentions in moral judgments (Chandler, Sokol & Wainryb, 2000; Fincham & Jaspars, 1979; Grueneich, 1982; Shiverick & Moore, 2007; Gweon, Dodell-Feder, Olson-Brown, Bedny & Saxe, in prep)

The right TPJ also undergoes subtle but systematic functional maturation in older children. Children aged 6 to 12 years have robust activity in the right TPJ when they listen to stories about people's thoughts and feelings, compared to stories about physical events. The region of activity in children is not smaller than that in adults; but it does show a striking difference in function. While right TPJ activity in adults is restricted to stories about thoughts, in children 6-9 years old the right TPJ responds equally to any stories about people (Kobayashi, Glover & Temple, 2007; Saxe et al., 2009). That is, by age 6 years the right TPJ is already selectively recruited by social information, but over the following 3-4 years, its function becomes even more specialised, within the domain of social information, to respond only when thinking about thoughts. Moreover, children's performance on sophisticated tests of ToM is correlated with a more selective response in their right TPJ, even after controlling for age (Gweon et al., in prep).

So cognitive development of ToM does seem to be related to maturation of the right TPJ, in both younger children (i.e. 4 year olds, measured with EEG) and older children (i.e. 5 -12 year olds, measured with fMRI). Furthermore, development of the right TPJ does not depend on visual access to other people, or on direct first-person experiences of the attributed mental states. Still, lots of questions remain to be answered. What aspects of developmental experience *are* necessary for ToM development? Would language-delayed children, who are delayed in ToM task performance, show delayed right TPJ maturation? Is right TPJ maturation also driven by biological factors? Which ones, and how?

The right TPJ and Autism Spectrum Disorders

Research on the maturation of the right TPJ, and ToM development, thus brings us back to the questions with which we began. Observation that children with ASD show a disproportionate impairment of ToM initially motivated the hypothesis that there might be a distinct neural substrate of ToM. Neuroscientific investigations have identified a plausible candidate for this

neural substrate: a region in the right TPJ whose function is linked to ToM in typically developing adults and children. To close the loop, all we need is evidence that the social impairments in ASD can be traced back to disfunction, or abnormal development, in the right TPJ. There is some evidence consistent with this hypothesis, but the mechanisms remain poorly understood.

In some studies, adults with ASD show reduced activation of brain regions implicated in ToM, including the right TPJ (Castelli et al., 2002; (Kana, Keller, Cherkassky, Minshew & Just, 2009), but in other studies adults with ASD show increased activation of the same regions (Mason, Williams, Kana, Minshew & Just, 2008), and still other studies find no difference (Happé et al., 1996). A rich interpretation of these conflicting results might treat both hyper- and hypo-activation as consistent evidence of abnormal function, whose manifestation depends on the specific task and context. However, a leaner interpretation is also possible: adults with ASD seem to be more heterogenous than the matched control population. Increased variability both within and between brains, in the ASD group, could lead to highly variable measurements, especially given the typically small sample size of fMRI studies (e.g. 8 - 16 participants per group).

One interesting new suggestion is that reliable differences between ASD and control groups may emerge not just in right TPJ activity, but in the connectivity between the right TPJ and other brain regions (Kennedy, Redcay & Courchesne, 2006). Connectivity between functional brain regions can be measured indirectly, by looking at the correlations between activity in different brain regions, independent of the task (Greicius, Krasnow, Reiss & Menon, 2003). Kana and colleagues (Kana et al., 2009) found that when adults with ASD watch animations of geometric shapes interacting (e.g. “flirting” or “coaxing), the activity in their right TPJ was less correlated with activity in medial frontal cortex, than in controls. A major goal for future research will be to study both activation, and connectivity, of the right TPJ in children with ASD, because the clearest signal of ASD in the brain may be not the endstate, but the trajectory of development (Courchesne, Carper & Akshoomoff, 2003; Redcay & Courchesne, 2005).

Conclusions

We've come a long way since the first cognitive neuroscience studies of ToM in 2000. Hundreds of papers have investigated the functions of the cortical regions those studies discovered. There are now whole journals devoted to social cognitive and social affective neuroscience. Looking back, I think the field has made remarkable progress. The view is even better, though, looking forward. So many questions about the cognitive and neural basis of ToM, in typical and a typical development, are waiting to be answered.

References

Amunts, A., Malikovic, M., Mohlberg, M., Schormann, S., & Zilles, Z. (2000). Brodmanns areas 17 and 18 brought into stereotaxic space-where and how variable?. *Neuroimage*, *11*.

Apperly, I. A., Samson, D., Carroll, N., Hussain, S., & Humphreys, G. (2006). Intact first- and second-order false belief reasoning in a patient with severely impaired grammar. *Soc Neurosci*, *1*(3-4), 334-48.

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *J Cogn Neurosci*, *16*(10), 1773-84.

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 Cogn Sci*, *9*(12), 572 - 577.

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory of mind development. *Dev Psychol*, *35*(5), 1311-20.

Avis, J., & Harris, P. L. (1991). Belief-Desire reasoning among baka children: Evidence for a universal conception of mind. *Child Dev*, *62*(3), 460-467.

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. *Journal of Child Psychology and Psychiatry*, *30*(2), 285-297.

Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a theory of mind?. *Cognition*, *21*, 37-46.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind.* Oxford: Oxford University Press.

Bedny, M., Pascual-Leone, A., & Saxe, R. R. (2009). Growing up blind does not change the neural bases of theory of mind. *Proc Natl Acad Sci U S A*, *106*(27), 11312-7.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-Selective areas in human auditory cortex. *Nature*, *403*(6767), 309-12.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *J Neurosci*, *17*(1), 353-62.

Binder, J. R., Swanson, S. J., Hammeke, T. A., Morris, G. L., Mueller, W. M., Fischer, M., et al. (1996). Determination of language dominance using functional MRI: A comparison with the wada test. *Neurology*, *46*(4), 978-84.

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating action understanding: Inferential processes versus action simulation. *Current Biology*, *17*(24), 2117 - 2121.

Brunet, E., Sarfati, Y., Hardy-Bayle, M. C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *Neuroimage*, *11*(2), 157-66.

Brunet, E., Sarfati, Y., Hardy-Baylé, M. C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *Neuroimage*, *11*(2), 157-66.

Bunge, S. A., Hazeltine, E., Scanlon, M. D., Rosen, A. C., & Gabrieli, J. D. E. (2002). Dissociable contributions of prefrontal and parietal cortices to response selection. *Neuroimage*, *17*(3), 1562 - 1571.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Dev*, *72*(4), 1032-53.

Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *J Exp Child Psychol*, *87*(4), 299-319.

Castelli, F., Frith, C., Happe, F., & Frith, U. (2002). Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, *125*(Pt 8), 1839-49.

Castelli, F., Happe, 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.

Chandler, M. J., Sokol, B. W., & Wainryb, C. (2000). Beliefs about truth and beliefs about rightness. *Child Development*, *71*(1), 91-97.

Charman, T., & Baron-Cohen, S. (1992). Understanding drawings and beliefs: A further test of the metarepresentation theory of autism: A research note. *J Child Psychol Psychiatry.*, *33*(6), 1105-12.

Corbetta, M., Shulman, G. L., Miezin, F. M., & Petersen, S. E. (1995). Superior parietal cortex activation during spatial attention shifts and visual feature conjunction. *Science*, *270*(5237), 802.

Courchesne, E. (1977). Event-Related brain potentials: Comparison between children and adults. *Science*, *197*(4303), 589-92.

Courchesne, E. (1978). Neurophysiological correlates of cognitive development: Changes in long-latency event-related potentials from childhood to adulthood. *Electroencephalogr Clin Neurophysiol*, *45*(4), 468-82.

Courchesne, E., Carper, R., & Akshoomoff, N. (2003). Evidence of brain overgrowth in the first year of life in autism. *JAMA*, *290*(3), 337-44.

Courchesne, E., Ganz, L., & Norcia, A. M. (1981). Event-Related brain potentials to human faces in infants. *Child Dev*, *52*(3), 804-11.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychol Sci*, *17*(12), 1082-9.

Decety, J., & Grezes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends Cogn Sci*, *3*(5), 172-178. 5_00001312 5_00001312.

Decety, J., Grèzes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., et al. (1997). Brain activity during observation of actions. Influence of action content and subject's strategy. *Brain*, *120 (Pt 10)*, 1763-77.

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, *298*(5600), 2013-5.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470-3.

Dunn, J., & Brophy, M. (2005). Communication, relationships, and individual differences in children's understanding of mind. In J. W. Astington, & J. A. Baird (Eds.), *Why language matters for theory of mind.* (pp. 50-69). Oxford: OUP.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E. J., & Shadlen, M. N. (1994). fMRI of human visual cortex. *Nature*, *369*(6481), 525.

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb Cortex*, *7*(2), 181-92.

Fincham, F., & Jaspars, J. (1979). Attribution of responsibility to the self and other in children and adults. *Journal of Personality and Social Psychology*, *37*(9), 1589-1602.

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

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., et al. (1995b). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*(2), 109-28.

Friston, K. J., Frith, C. D., Fiddle, P. F., & Frackowiak, R. S. J. (1993). Functional connectivity: The principal-component analysis of large (PET) data sets. *Journal of Cerebral Blood I'low Or. D Ucicbolmr*, *3*, 5-14.

Frith, U. (1997). The neurocognitive basis of autism. *Trends Cogn Sci*, *1*(2), 73-77.

Frith, C. D., & Frith, U. (1999). Interacting minds--a biological basis. *Science*, *286*(5445), 1692-5.

Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current Directions in Psychological Science*, *10.*

Gallagher, H. L., Happe, 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.

German, T. P., Niehaus, J. L., Roarty, M. P., Giesbrecht, B., & Miller, M. B. (2004). Neural correlates of detecting pretense: Automatic engagement of the intentional stance under covert conditions. *J Cogn Neurosci*, *16*(10), 1805-17.

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *J Cogn Neurosci*, *19*(11), 1803-14.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *Neuroreport*, *6*(13), 1741-6.

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proc Natl Acad Sci U S A*, *100*(1), 253-8.

Grill-Spector, K., & Malach, R. (2001). Fmr-Adaptation: A tool for studying the functional properties of human cortical neurons. *Acta Psychol (Amst)*, *107*(1-3), 293-321.

Grill-Spector, K., Sayres, R., & Ress, D. (2006). High-Resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nat Neurosci*, *9*(9), 1177-85.

Grueneich, R. (1982). The development of children's integration rules for making moral judgments. *Child Development*, 887-894.

Gweon, H., Dodell-Feder, D., Olson-Brown, E., Bedny, M., & Saxe, R. (in prep). *Developmental change in the neural mechanisms for theory of mind.*

de Haan, M., Johnson, M. H., & Halit, H. (2003). Development of face-sensitive event-related potentials during infancy: A review. *Int J Psychophysiol*, *51*(1), 45-58.

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Dev Sci*, *6*(3), 346-59.

Hamilton, A. F., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. *J Neurosci*, *26*(4), 1133-7.

Happé, F., Ehlers, S., Fletcher, P., Frith, U., Johansson, M., Gillberg, C., et al. (1996). 'Theory of mind' in the brain. Evidence from a PET scan study of asperger syndrome. *Neuroreport*, *8*(1), 197-201.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425-30.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biol Psychiatry*, *51*(1), 59-67.

Hinds, O. P., Rajendran, N., Polimeni, J. R., Augustinack, J. C., Wiggins, G., Wald, L. L., et al. (2008). Accurate prediction of V1 location from cortical folds in a surface coordinate system. *Neuroimage*, *39*(4), 1585-99.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nat Neurosci*, *3*(1), 80-4.

Holcomb, P. J., Coffey, S. A., & Neville, H. J. (1992). Visual and auditory sentence processing: A developmental analysis using event-related brain potentials. *Developmental Neuropsychology*, *8*(2-3), 203-241.

Hyde, D. C., Jones, B. L., Porter, C. L., & Flom, R. (2010). Visual stimulation enhances auditory processing in 3-month-old infants and adults. *Dev Psychobiol*, *52*(2), 181-9.

Kana, R. K., Keller, T. A., Cherkassky, V. L., Minshew, N. J., & Just, M. A. (2009). Atypical frontal-posterior synchronization of theory of mind regions in autism during mental state attribution. *Soc Neurosci*, *4*(2), 135-52.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J Neurosci*, *17*(11), 4302-11.

Kennedy, D. P., Redcay, E., & Courchesne, E. (2006). Failing to deactivate: Resting functional abnormalities in autism. *Proc Natl Acad Sci U S A*, *103*(21), 8275-80.

Kobayashi, C., Glover, G. H., & Temple, E. (2007). Children's and adults' neural bases of verbal and nonverbal 'theory of mind'. *Neuropsychologia*, *45*(7), 1522-32.

Lee, K., Olson, D. R., & Torrance, N. (1999). Chinese children's understanding of false beliefs: The role of language. *J Child Lang*, *26*(1), 1-21.

Leslie, A., & Thaiss, L. (1992). Domain specificity in conceptual development. *Cognition*, *43*, 225-51.

Lewis, C., & Osborne, A. (1990). Three-Year-Olds' problems with false belief: Conceptual deficit or linguistic artifact?. *Child Dev*, *61*(5), 1514-9.

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.

Liu, D., Sabbagh, M. A., Gehring, W. J., & Wellman, H. M. (2009). Neural correlates of children's theory of mind development. *Child Development*, *80*(2), 318-26.

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Dev Psychol*, *44*(2), 523-31.

Luck, S. J. ,. 1. (2005). *An introduction to the event-related potential technique* (illustrated ed.). Cambridge, Mass. : MIT Press.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835.

Mason, R. A., Williams, D. L., Kana, R. K., Minshew, N., & Just, M. A. (2008). Theory of mind disruption and recruitment of the right hemisphere during narrative comprehension in autism. *Neuropsychologia*, *46*(1), 269-80.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *Neuroimage*, *28*(4), 757-62.

Mitchell, P., & Lacohée, H. (1991). Children's early understanding of false belief. *Cognition*, *39*(2), 107-27.

Newton, A. M., & de Villiers, J. G. (2007). Thinking while talking. *Psychological Science*, *18*(7), 574.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?. *Science*, *308*(5719), 255-8.

Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: The perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *J Cogn Neurosci*, *16*(10), 1706-16.

Perner, J. (1991). Understanding the representational mind. . *Cambridge, MA: 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. *Soc Neurosci*, *1*(3-4), 245-58.

Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development*, *76*(2), 502-517.

Peterson, C. P., & Siegal, M. S. (1999). Representing inner worlds: Theory of mind in autistic, deaf, and normal hearing children. *Psychological Science*, *10*(2), 126-129.

Puce, A., Constable, R. T., Luby, M. L., McCarthy, G., Nobre, A. C., Spencer, D. D., et al. (1995). Functional magnetic resonance imaging of sensory and motor cortex: Comparison with electrophysiological localization. *J Neurosurg*, *83*(2), 262-70.

Pujol, J., Deus, J., Losilla, J. M., & Capdevila, A. (1999). Cerebral lateralization of language in normal left-handed people studied by functional MRI. *Neurology*, *52*(5), 1038-43.

Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science: A Journal of the American Psychological Society/Aps.*

Redcay, E., & Courchesne, E. (2005). When is the brain enlarged in autism? A meta-analysis of all brain size reports. *Biol Psychiatry*, *58*(1), 1-9.

Redcay, E., Haist, F., & Courchesne, E. (2008). Functional neuroimaging of speech perception during a pivotal period in language acquisition. *Dev Sci*, *11*(2), 237-52.

Redcay, E., Kennedy, D. P., & Courchesne, E. (2007). Fmri during natural sleep as a method to study brain function during early childhood. *Neuroimage*, *38*(4), 696-707.

Roth, D., & Leslie, A. M. (1998). Solving belief problems: Toward a task analysis. *Cognition*, *66*(1), 1-31.

Rowe, J. B., Toni, I., Josephs, O., Frackowiak, R. S. J., & Passingham, R. E. (2000). The prefrontal cortex: Response selection or maintenance within working memory?. *Science*, *288*(5471), 1656.

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *J Exp Child Psychol*, *80*(3), 201-24.

Ruffman, T., Slade, L., Rowlandson, K., Rumsey, C., & Garnham, A. (2003). How language relates to belief, desire and emotion understanding. *Cognitive Development*, *113*, 1-20.

Sabbagh, M. A., Fen, X., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind. *Psychological Science*, *17*(1), 74-81.

Sabbagh, M. A., & Taylor, M. (2000). Neural correlates of theory-of-mind reasoning: An event-related potentialstudy. *Psychological Science*, *11*.

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain*, *128*(Pt 5), 1102-11.

Saxe, R. 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.

Saxe, R. (2006). Uniquely human social cognition. *Curr Opin Neurobiol*, *16*(2), 235-9.

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., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, *17*(8), 692-699.

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., Schulz, S., & Jiang, Y. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Soc Neurosci*, *1*(3/4), 284 - 298.

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.

Shiverick, S. M., & Moore, C. F. (2007). Second-Order beliefs about intention and children's attributions of sociomoral judgment. *J Exp Child Psychol*, *97*(1), 44-60.

Siegal, M., & Varley, R. (2006). Aphasia, language, and theory of mind. *Soc Neurosci*, *1*(3-4), 167-74.

Slaughter, V. (1998). Children's understanding of pictorial and mental representations. *Child Dev*, *69*(2), 321-32.

Smith, A. T., Greenlee, M. W., Singh, K. D., Kraemer, F. M., & Hennig, J. (1998). The processing of first- and second-order motion in human visual cortex assessed by functional magnetic resonance imaging (fmri). *J Neurosci*, *18*(10), 3816-30.

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

Stuss, D. T., & Benson, D. F. (1984). Neuropsychological studies of the frontal lobes. *Psychological Bulletin*, *95*(1), 3-28.

Tootell, R. B., Reppas, J. B., Kwong, K. K., Malach, R., Born, R. T., Brady, T. J., et al. (1995). Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *J Neurosci*, *15*(4), 3215-30.

de Villiers, J. (2000). Language and theory of mind: What are the developmental relationships? In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds.* Oxford: Oxford University Press.

de Villiers, J., & de Villiers, P. A. (2000). Linguistic determinism and the understanding of false beliefs. In P. Mitchell, & K. J. Riggs (Eds.), *Children's reasoning and the mind.* (pp. 191-228). Hove, England: Psychology Press/ Taylor & Francis (UK).

de Villiers, P. A. (2005). The role of language in theory-of-mind development: What deaf children tell us. In J. W. Astington, & J. A. Baird (Eds.), *Why language matters for theory of mind.* (pp. 266-97). Oxford: OUP.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., et al. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *Neuroimage*, *14*(1 Pt 1), 170-81.

Wellman, H., & Bartsch, K. (1998). Young children's reasoning about beliefs. *Cognition*, *30*(3), 239-277.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Dev*, *75*(2), 523-41.

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

Yazdi, A. A., German, T. P., Defeyter, M. A., & Siegal, M. (2006). Competence and performance in belief-desire reasoning across two cultures: The truth, the whole truth and nothing but the truth about false belief?. *Cognition*, *100*(2), 343-68.

Young, L., Dodell-Feder, D., & Saxe, R. (In press). What gets the attention of the temporo-parietal junction? An fmri investigation of attention and theory of mind. *Neuropsychologia.*

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. *J Cogn Neurosci*, *21*(7), 1396-405.

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

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proc Natl Acad Sci U S A*, *104*(20), 8235-40.

Zatichik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and false photographs. *Cognition*, *35*(1), 41-68.