Theory of Mind: How brains think about thoughts

Rebecca Saxe and Liane Young

Department of Brain and Cognitive Sciences, MIT

At the heart of comedy and tragedy, there is often a false belief. Titania doesn't know she's in love with a donkey. Romeo thinks Juliet is dead. Human audiences are brought to laughter and tears. Imagine, though, an audience that doesn't have a concept of belief, that cannot think about other people's thoughts at all. These plots would make no sense. In fact, the whole notion of theatre, of watching actors depict a fictional story, could never get off the ground.

Our minds and brains have, among their most astonishing capacities, the ability to see behind people's physical actions to their internal causes, thoughts and intentions. That is, we have a Theory of Mind (ToM) for understanding and interpreting the external actions of others. When the audience thinks "Romeo doesn't know that Juliet wants her parents to think that she is dead", that thought consists of a pattern of firing across a group of neurons somewhere in each person's brain. This fact is both obvious (what is the alternative?) and mind-boggling. How are those neurons doing it?

To get the answers, we need to be able to study the human brain in action. Unlike the traditional neuroscience topics covered in this volume - perception, motor control, attention, memory, and emotion - uniquely human cognitive capacities, like language and social cognition, cannot be studied in the brains of non-human animals. The invention of functional neuroimaging has therefore opened up many topics that,

historically, belonged only to social sciences: how we think about people, how we think about thoughts, how we make moral judgments, and more.

Although the neuroscience of ToM is only around a decade old, we will review evidence that begins to address some fundamental questions. What are the neural substrates of ToM? Are there distinct brain regions selectively recruited for ToM (as there are regions for vision, audition, motor control, etc.)? If so, what are (and aren't) these brain regions doing? Are there distinct cognitive components of ToM? Answers to these questions provide the foundation for a cognitive neuroscience of Theory of Mind.

Where in the brain do people think about thoughts?

Human adults can think about other people as having an infinite array of beliefs and desires, ranging from trivial to sublime, from familiar to exotic, from simple to remarkably complex.

For example, consider the following story: Sally and Anne go to the same high school. Sally doesn't suspect that Anne knows that Sally's boyfriend Tom believes that the tooth fairy stole the quarterback's lucky tooth before the big game, jinxing the team. Anne also knows that Tom will propose to Sally at graduation, so Anne realises that only she can stop their engagement.

Even though this story is highly complex, the people are unfamiliar to you, and you likely have never considered the possibility of the tooth fairy's interference in a football game, you can nevertheless make sense of this story, and predict and explain the characters' actions and emotions. How do you do it? What is happening in your brain

while you read the story? Let's imagine following the story from the pattern on the page to the pattern in your brain.

First, the pattern of light and dark on the page reaches your eyes, and then your visual cortex. Here the brain begins to recognise shapes, and to test hypotheses about which letters and words are on the page. Soon, language brain regions are involved, helping to transform the representations from orthographic symbols to words and sentences that describe objects, events, and ideas - these representations are complex. As you build up a mental representation of all the elements in the story, your working memory helps to hold and manipulate the elements, while executive control supports shifts between the competing components of the event. In particular, executive control helps you keep track of what really happened, distinct from what Sally didn't suspect that Anne knew that Tom believed was happening. As you begin to understand and represent the events of the story, specific aspects of the story become clear. This is a story about people, social relationships, and human actions. This story requires you to think about different perspectives or representations of the same facts; that is, it requires the capacity to form "meta-representations". And this story requires you to think about people's thoughts, beliefs, desires, motivations, and emotions.

Remarkably, human cognitive neuroscience can already help us pinpoint where in the brain each one of these different cognitive processes is occurring. Other chapters of this handbook describe the brain regions and processes involved in vision, word recognition, language comprehension, working memory, and executive function. Most interesting for our current purposes are three cognitive processes (and associated brain regions) that appear to be disproportionately necessary for reading and understanding a story about people and what they are thinking: (a) representing people and social

relations (e.g., dorsal medial prefrontal cortex); (b) representing representations (e.g., left temporo-parietal junction); and (c) representing *mental* representations (e.g. right temporo-parietal junction), that is, thinking about thoughts.

All of these brain regions had a high metabolic response while you were reading the story about Sally, Anne and Tom, but for different reasons - these brain regions perform different functions in helping you to perceive and reason about the story. To understand how we infer these different functions, it's helpful to imagine an (implausible) meta-experiment, in which we could present participants with 5 different kinds of stimuli and see which patterns of responses we observe, and where. Each brain region or system would reveal different patterns of functional response across the categories (see Figure 1 for a schematic representation of the imaginary experiment, and Figure 2 for sample stimuli from actual experiments).

For example, there is a region near the calcarine sulcus that responds robustly when people read stories and look at pictures but not when people listen to stories or to music. Meanwhile, there is no difference in this region's response to the specific content of the stories, i.e. whether the stories focus on physical objects, temporal changes, people, or their thoughts. However, the response in this brain region to the same story is very different depending on whether the story is presented visually (a high response) or aurally (a low response). Correspondingly, people with damage near the calcarine sulcus cannot perceive visually presented pictures or sentences but have no trouble understanding aural language or thinking about thoughts. Based on this pattern, we can diagnose that the cortex near the calcarine sulcus contains a brain region that is involved in visual perception (ref to vision chapter). This, of course, would not be news.

The visual system is one of the best-understood parts of the brain; none of the other brain regions we will consider here is affected by the modality of the stimulus.

Relying on a similar logic, we can look for patterns of functional responses and selective deficits, to infer the cognitive functions of other less well-understood brain regions, and also to learn about how these cognitive functions are related in the brain. For example, there is a brain region in the left dorsolateral prefrontal cortex (left DLPFC) that shows a high response for stimuli requiring difficult reasoning, especially for balancing competing ideas or responses. This brain region shows a high response when people read a story that describes two competing versions of reality: one past and one present, or one in a photograph and one in reality, or one that someone believes and one that actually happened. This brain region also shows a high response when you try to name the red ink colour of the word "green", compared to the blue ink colour of the word "blue" - the standard Stroop task manipulation of conflict (MacDonald, Cohen, Stenger, & Carter, 2000). Damage to this brain region therefore makes it difficult to resolve such cognitive competition, and as a result can make it difficult for people to reason accurately about another person's thoughts and beliefs in certain cases. For example, patients with left DLPFC damage wouldn't be able to balance their own ideas about Tom and the competing ideas about Tom held by Anne and Sally. Instead, these patients would just stick with their own perspective: if Tom is crazy, then Sally won't want to marry him. On the other hand, if there is no conflict in the story - for example, when we hear that Anne thinks only she can stop the engagement, which might be true or false and doesn't conflict with any other ideas - these patients have no problems thinking about beliefs per se, and predicting what Anne will do next (Apperly, Samson, Chiavarino, & Humphreys, 2004).

A brain region in the left temporo-parietal junction (TPJ) shows a second functional profile. The left TPJ response is high for any story, picture, or task that requires reasoning about perspectives, or representations of the world - whether those representations are mental representations (like people's beliefs about the world), or physical representations (like photographs of the world). Correspondingly, patients with damage to the left TPJ have difficult with tasks that require reasoning about beliefs, photographs and maps, but not with other 'high-conflict' tasks, like naming the ink colour of the word "green", printed in red ink. These patients have trouble thinking about any kind of belief or indeed any representation at all, including a physical representation like a photograph, whether or not these representations conflict with reality. So we can infer that the left TPJ is involved in meta-representation, including but not limited to representing *mental* representations.

The functions of the DLPFC and the LTPJ may seem similar, but they have been elegantly dissociated by Dana Samson, Ian Apperly and colleagues, in studies of patients with selective lesions. To get a sense for the dissociation, imagine the story continues on, to reveal who *actually* stole the lucky tooth: a crazy ex-girlfriend of the quarterback. Now, if you must answer, "what does Tom think happened to the quarter-back's tooth?", you might consider three possible answers. First, the correct answer, which depends on keeping track of Tom's false belief, would be "he thinks the tooth fairy stole it." Second, if you couldn't hold on to Tom's belief in the face of the stronger competition from your knowledge of what really happened, then the 'reality-error' answer would be "he thinks an ex-girlfriend stole it". This is the kind of error produced by DLPFC damage. Third, though, if you could resist competition from reality, but couldn't hold on to a representation of Tom's belief, then you might just seek a likely

explanation for a quarterback's missing tooth, and make the 'appearance-error': "he thinks it was knocked out during a game." Left TPJ damage, but not DLFC damage, leads to 'appearance' errors (Samson, Apperly, Chiavarino, & Humphreys, 2004).

In sum, thinking about thoughts depends on many cognitive functions that are not specific to ToM. ToM tasks are often hard logical problems, involving complex reasoning and perspective shifts, and therefore rely on multiple brain such regions - DLPFC and LTPJ are only examples. In addition, though, human cognitive neuroscience has revealed another group of brain regions, with a notably different pattern of response: these regions are involved specifically in thinking about other people.

Returning to our imaginary experiment, a third functional profile can be found in a the medial prefrontal cortex (MPFC)[1]. Here we would not see a high response to stories about photographs, or physical interactions, or temporal changes; only stories with people and social relationships elicit a response in the MPFC. Thus, we can infer that the MPFC is involved specifically in social cognition.

Finally, a brain region near the right temporo-parietal junction (RTPJ) shows a robust response during our original story (regardless of modality), but does not respond to any of the other conditions in this imaginary experiment - not to difficult logical problems, or stories about photographs, or stories about people and social relationships (R. Saxe & Kanwisher, 2003; R. Saxe & Powell, 2006). Of the conditions in our imaginary experiment, the RTPJ region shows a high response only when the story describes someone's thoughts and beliefs.

[1] Here, we describe the MPFC as a single region, though research has shown dissociable sub-regions within the MPFC, including the ventral MPFC and the dorsal MPFC. In some cases, these sub-regions have importantly different response profiles (Mitchell, Macrae, & Banaji, 2006). Here we try to focus on features of the response that are common across sub-divisions of the MPFC, for simplicity, but we strongly urge readers specifically interested in the MPFC to consider these differences, as described in other reviews (e.g. (Adolphs, 2009)).

Regions in MPFC and RTPJ are most commonly recruited together, possibly because thinking about thoughts usually also involves thinking about people and social relationships (broadly construed; see Figure 2). However, careful experiments reveal fascinating functional dissociations between these two regions. For example, activity in your RTPJ was high when you read about Sally, Anne and Tom's true and false beliefs, but would be low if you were reading instead about what Sally looks like (i.e. her physical traits) and whether she is stubborn or lazy (i.e. her personality traits), where Anne comes from and how many siblings she has (i.e. her history and status), or what Tom prefers to eat for breakfast (i.e. his stable preferences). Even reading about how Sally feels when she's hungry or tired or in physical pain would not elicit a robust response in the RTPJ (Bedny, Pascual-Leone, Dodell-Feder, Fedorenko, & Saxe; Jenkins & Mitchell, 2009; R. Saxe & Powell, 2006 ; R. R. Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009). Regions in the MPFC, on the other hand, would show high activity for most of this information, especially descriptions of stable preferences and personality traits (Jenkins & Mitchell, 2009). One factor that matters to the response in MPFC, but not in RTPJ, is the person being described. The response in the MPFC region would be much higher if Sally, Anne and Tom were friends of yours - either people you found similar to yourself, or people who were emotionally close to you (Krienen, Tu, & Buckner; Mitchell, et al., 2006). By contrast, the RTPJ does not seem to care about the identity of the target.

What happens when these regions, the MPFC and the RTPJ, are not functioning properly? Damage to MPFC often leads to problems, for example, for thinking about other people's emotions (Shamay-Tsoory, Tomer, Berger, & Aharon-Peretz, 2003), but not necessarily for thinking about people's thoughts (Bird, Castelli, Malik, Frith, &

Husain, 2004). Selective damage to the RTPJ has not been as well studied, but similar evidence comes from an experiment in which we can produce a temporary or reversible 'lesion', using a tool called transcranial magnetic stimulation (TMS). To understand the experiment, it will help to start with a new story about Sally and Anne.

Imagine Sally is making dinner for Anne. Based on something Anne said, Sally believes that Anne is violently allergic to peanuts. Sally grinds up some peanuts, and mixes them into the soup, which she then serves to Anne. In fact, Anne is allergic to coconuts but not peanuts, so she happily enjoys the soup. Now, did Sally do anything morally wrong? From the outside, nothing bad happened. Sally served Anne some delicious soup. Most people, though, say that what Sally did was very wrong, because Sally *believed* she was doing something wrong. The opposite case presents an even starker contrast. Imagine Sally adds coconut shavings to the soup, but she has absolutely no idea that Anne is allergic to coconuts or anything else. Now, Anne eats the soup and becomes fatally ill. Did Sally do anything morally wrong? In spite of the tragic consequences of her actions, most people say that what Sally did was not very wrong - because she reasonably believed her actions would not hurt anyone. These scenarios provide a sensitive measure of how much people are thinking about thoughts. The more you think about thoughts, the more you will blame Sally for attempting (but failing) to poison Anne, and the more you will forgive her for accidentally making Anne sick (and the more active your RTPJ will be! (Young & Saxe, 2009)).

To test the role of the RTPJ in thinking about thoughts, we briefly disrupted normal neural function specifically in the RTPJ, using fMRI-guided TMS. When the RTPJ has been targeted with TMS, moral judgments shift. Innocent accidents appear more blameworthy, while failed attempts appear less blameworthy, as though it matters less

what Sally *believed* she was doing, and it matters more what she actually does (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). (People don't lose the ability to make moral judgments altogether; they still say it's completely morally wrong to intentionally kill, and not wrong at all to simply serve someone soup). These results fit very nicely with the fMRI studies. Activity in the RTPJ is correlated across time, across people, and across individual stories, specifically with the need to think about thoughts (Young & Saxe, 2008, 2009; Bruneau & Saxe, unpublished data), when function in the RTPJ is disrupted, people think less about thoughts and more about other features of the stories.

Understanding the neural basis of theory of mind will therefore probably begin with understanding the function(s) of these regions, that is, the MPFC, for thinking about people, and the RTPJ, for thinking about thoughts, along with the interactions between these regions with each another and with the rest of the brain. Provisionally, though, there seem to be patches of cortex in the human brain whose functions are specifically related to ToM (RTPJ) or social cognition (MPFC). This claim raises key questions that we address in the next section: What does it mean to say that a brain region's function is 'specifically related to ToM'? What are and aren't these brain regions doing?

How does the brain think about thoughts?

From a certain perspective, ToM is a miracle. After all, thoughts are invisible: no one has ever had any direct evidence of another person's mental experience. How do our brains cross the gulf between our minds? One idea that may demystify the leap is that we understand other minds by "simulation" (Goldman, 2006). The central idea of simulation is that we understand other people because they are similar to us: they execute similar movements, and experience similar sensations, and make similar

decisions, using a body and mind similar to our own. As a result, we could use our own mind (and body) as an analogue for another person's mind. We could recreate in ourselves a copy of their actions and sensations, and recapitulate our own experiences in order to understand theirs. Could *this* be the distinctive function of ToM brain regions: to construct appropriate and useful simulations of other minds?

People do seem to simulate the actions they observe, by activating matching motor representations in their own brain and body. When a person watches someone else act, the observer can't help but activate the same muscles and motor plans for that action (Fadiga, Craighero, & Olivier, 2005). As a result, action observation interferes with action execution, and action execution interferes with action observation (Zwickel, Grosjean, & Prinz, 2010a, 2010b). Even when the other person's actions are invisible, simply knowing about someone else's incompatible action can cause interference. In an elegant series of studies, Saebanz and colleagues showed that interference from thinking about *another* person's actions is comparable to competition from *one's own* actions (Sebanz, Knoblich, & Prinz, 2003). That is, if you are trying to push the left button, but thinking about pushing the right button, these two motor plans interfere with each other and slow you down. Amazingly, thinking about someone else's action has the same effect: when you know someone else is supposed to push the right button, you yourself are slower to push the left button! A similar pattern occurs when you observe what other people see. Seeing that another person sees more or less than you do can actually impair your ability to report what you yourself are seeing, as though you automatically compute the other person's view, which then competes with your own view for verbal report (Samson, Apperly, Braithwaite, Andrews, & Scott, 2010). These

results show that watching and understanding another person's action compete for the same cognitive and neural resources as executing one's own action.

Neural evidence converges on the same simulation story. Activity in the parietal cortex while watching someone else perform a simple hand action is suppressed if the participant had just previously made the same hand action, suggesting that the representation of one's own action can be partially 'recycled' during observation of someone else's (Chong, Cunnington, Williams, Kanwisher, & Mattingley, 2008). And, complementarily, watching someone else's hand movements leads to sub-threshold preparatory activity in one's own motor cortex and hand muscles: this activity can be seen if it is artificially pushed over the threshold by a pulse of transcranial magnetic stimulation (Sturmer, Siggelkow, Dengler, & Leuthold, 2000). Furthermore, these activations seem to be modulated by experience: the more experience the observer has had actually performing a particular action, the more his or her motor cortex is activated while observing others performing the same action. In one elegant example, the motor cortex of ballet dancers showed more activity when dancers observed gender-specific movements that they themselves had more experience executing (Calvo-Merino, Glaser, Grezes, Passingham, & Haggard, 2005; Cross, Hamilton, & Grafton, 2006), but equal experience observing, in dancers of both genders.

A similar pattern holds for observing physical sensations in another person, especially physical pain. A common group of brain regions are recruited when people feel their own pain, and when they see someone else in pain. Experiencing pain leads to brain activity in the "pain matrix", including regions in cingulate cortex, secondary sensory cortex, and bilateral insula. When observers witness other people in physical pain, some of the same brain regions are activated (Botvinick, et al., 2005; Jackson,

Rainville, & Decety, 2006; Singer & Lamm, 2009; Singer, et al., 2004). Activity in some of these regions is correlated with the intensity of pain, either experienced (Peyron, Laurent, & Garcia-Larrea, 2000) or attributed (Saarela, et al., 2007).

In sum, we appear to 'simulate' other people's actions and experience: as observers, we recruit (some of) the same representations as the target. Simulations - the re-cycling of similar representations between first-person experience and third-person attributions - thus seem to reflect a general principle of how we bridge the gap between two separate human minds. Is activity in the RTPJ and MPFC also modulated by whether the mental states we attribute to other people are similar to mental states we've experienced in the first person? Similar to the logic of 'simulation' for actions and experiences, do we understand someone else's desire to become a neurosurgeon, or belief that the Red Sox will win the World Series, by activating the same representations in our own mind as if we ourselves had that desire, or held that belief?

As we described above, regions in the MPFC are modulated by a related issue: whether the *target person* is, overall, similar or close to oneself. For example, MPFC is recruited when you are asked about the personality, preferences, and habits of people who are similar and/or emotionally important to you, like your mother, compared to when asked about people who are dissimilar or less close, like President Obama (Mitchell, et al., 2006). There even seems to be some 'shared representation' of your own preferences and traits, and those of similar others. If you have just been thinking about your own preferences, and then transfer to thinking about the preferences of a similar other, the response in the MPFC is 'adapted' (i.e. relatively low), suggesting the two processes depend on shared neural substrates (Jenkins, Macrae, & Mitchell, 2008). When put to the test, though, the MPFC response does not depend on similarity (or first

person experience), but on emotional closeness. The MPFC response is higher for emotionally close friends who are not similar to oneself, than for strangers who are very similar (Krienen, Tu, & Buckner, 2010). Unlike the motor representations of ballet dancers, which really do depend on first person experience, the response in MPFC during personality trait attribution reflects an assessment of social or personal significance.

The key region, though, for representing others' thoughts is the RTPJ. Here too the evidence against 'simulation' of other minds is clear. The RTPJ does not recapitulate the observer's own analogous thoughts and experiences, but is recruited for thinking about other people's thoughts even when those thoughts are maximally different from one's own.

Initially (R. Saxe & Wexler, 2005), we manipulated our participants' experience with specific beliefs and desires by generating examples of beliefs and desires unlikely to be frequently held by our participants (MIT undergraduates): a belief that conflicts are best resolved by physical violence, or a desire for one's partner to have an affair. Indeed, a post-scan survey confirmed that our participants found these beliefs and desires unfamiliar. Nevertheless, the RTPJ did not show less (or more) activation when reading about culturally-distant beliefs and desires, compared to more familiar counterparts. First-person experience holding a particular mental state did not seem to affect neural activation when people attributed that state to somebody else. Instead, activation in RTPJ was modulated by a different factor: whether the specific belief or desire made sense, given the background and culture of the target person. Beliefs about violence are more expected in members of a gang; acceptance of an affair fits with a person who has joined a cult. More generally, we expect other people to be coherent, unified entities,

and we strive to resolve inconsistencies with that expectation (Hamilton & Sherman, 1996): when someone's behaviour violates our previous impression of that person, we spend more time searching for the behaviour's causes (Hamilton 1988). Likewise, the response in the RTPJ was modulated by whether a character's beliefs and desires were congruent with other information about that person. That is, the RTPJ appeared to reflect a process of constructing a coherent model of the other person's mind, without reference to the participant's own mental states.

Later, we replicated this basic result with a different strategy. Instead of culturally unfamiliar beliefs, we asked participants to attribute common-sense beliefs ("John believes that swimming is a good way to cool off") or absurd beliefs ("John believes that swimming is a good way to grow fins"; (Young, Dodell-Feder, & Saxe, 2010)). Again, activity in RTPJ was no higher for attributing common-sense versus absurd beliefs.

In the third experiment (Bedny, Pascual-Leone, & Saxe, 2009), we pushed the prediction even further: we asked people to attribute to other people a mental state that they themselves could never experience. To do this, we asked individuals who had been blind since birth to reason about experiences of hearing (which are very familiar) and seeing (which they could never experience themselves but frequently hear others describing). We found that first-person experience of seeing is not necessary for the development of normal neural representations of another person's experiences of seeing. The RTPJ was recruited similarly for reasoning about beliefs formed based on seeing and based on hearing, in both sighted and blind adults. Apparently, recapitulating a similar first person experience is not necessary for the normal representation of someone else's experience.

In sum, thinking about thoughts does not show the same functional profile as observing actions or experiences. Activity in the key brain regions, the MPFC and especially the RTPJ, is not affected by people's first person experience or how similar the beliefs and desires are to their own beliefs and desires[2]. This is part of what makes humans' theory of mind so powerful: we can understand, explain, predict and judge other people's actions, even when they depend on beliefs and desires that we don't share and indeed have never experienced. We can imagine how Tom will act, given he believes in the tooth fairy, and what Anne will do to prevent Sally from marrying him, without knowing the people or giving any actual credence to their beliefs. That's part of what makes watching tragedy and comedy so gripping, and the human actions that unfold in them so predictable.

Conclusions

This chapter summarises the data that provide a foundation for a future neuroscience of Theory of Mind. Although there has been a furious burst of activity, studying the neural basis of ToM, in the last ten years, and hundreds of papers have been published, the most important questions remain unanswered. We have provided some evidence, for example, that the RTPJ and MPFC are not involved in 'simulating' other people's minds, based on the observer's own first person experience with similar beliefs and desires. So what computations *are* these brain regions doing? We have

[2] There is also a key conceptual difference between the studies of action 'simulation' and studies of theory of mind, beyond the empirical differences we've described here. While observing actions, there is activity in the same brain regions that are used during action execution - actually making body movements. On a strict analogy, simulation should predict that we understand beliefs and desires using the same brain regions we use for *having* beliefs and desires; and we think about other people's personalities using the same brain regions that we use for *having* our own personality. That is, we would recognize laziness in others using the brain regions that we use for being lazy. But upon reflection, this prediction doesn't make sense. There can't be specific brain regions for having a personality or having a belief; personalities and beliefs aren't specific cognitive processes or representations, but summary descriptions of behavioral tendencies. By contrast, the RTPJ and MPFC are associated with attributing thoughts and personality traits, which do require specific cognitive processes and representations.

described evidence that the RTPJ doesn't distinguish between true versus false beliefs, or hard versus easy inferences about beliefs. So which features of beliefs and desires *does* the RTPJ represent, and how? Finally, we don't know how, or why, human adults come to have brain regions specifically involved in thinking about people and their thoughts. What are the homologues of RTPJ and MPFC in other animals, and what are their functions? When do these regions mature in the course of human childhood, and why? All of these questions are on the table for the next decade of the neuroscience of Theory of Mind.

References

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annu Rev Psychol, 60*, 693-716.

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), 1773-1784.

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. Language processing in the occipital cortex of congenitally blind adults. *Proc Natl Acad Sci U S A, 108*(11), 4429-4434.

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), 11312-11317.

Bird, C. M., Castelli, F., Malik, O., Frith, U., & Husain, M. (2004). The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. *Brain, 127*(Pt 4), 914-928.

Botvinick, M., Jha, A. P., Bylsma, L. M., Fabian, S. A., Solomon, P. E., & Prkachin, K. M. (2005). Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. *Neuroimage, 25*(1), 312-319.

Calvo-Merino, B., Glaser, D. E., Grezes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: an FMRI study with expert dancers. *Cereb Cortex, 15*(8), 1243-1249.

Chong, T. T., Cunnington, R., Williams, M. A., Kanwisher, N., & Mattingley, J. B. (2008). fMRI adaptation reveals mirror neurons in human inferior parietal cortex. *Curr Biol, 18*(20), 1576-1580.

Cross, E. S., Hamilton, A. F., & Grafton, S. T. (2006). Building a motor simulation de novo: observation of dance by dancers. *Neuroimage, 31*(3), 1257-1267.

Fadiga, L., Craighero, L., & Olivier, E. (2005). Human motor cortex excitability during the perception of others' action. *Curr Opin Neurobiol, 15*(2), 213-218.
Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. New York: Oxford University Press.
Hamilton, D., & Sherman, S. (1996). Perceiving Persons and Groups. *Psychological Review, 103*(2), 336-355.
Jackson, P. L., Rainville, P., & Decety, J. (2006). To what extent do we share the pain of others? Insight from the neural bases of pain empathy. *Pain, 125*(1-2), 5-9.
Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proc Natl Acad Sci U S A, 105*(11), 4507-4512.
Jenkins, A. C., & Mitchell, J. P. Medial prefrontal cortex subserves diverse forms of self-reflection. *Soc Neurosci*, 1-8.
Jenkins, A. C., & Mitchell, J. P. (2009). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cereb Cortex, 20*(2), 404-410.
Krienen, F. M., Tu, P. C., & Buckner, R. L. Clan mentality: evidence that the medial prefrontal cortex responds to close others. *J Neurosci, 30*(41), 13906-13915.
Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: evidence that the medial prefrontal cortex responds to close others. *J Neurosci, 30*(41), 13906-13915.
MacDonald, A. W., 3rd, 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), 1835-1838.
Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*(4), 655-663.
Peyron, R., Laurent, B., & Garcia-Larrea, L. (2000). Functional imaging of brain responses to pain. A review and meta-analysis (2000). *Neurophysiol Clin, 30*(5), 263-288.
Saarela, M. V., Hlushchuk, Y., Williams, A. C., Schurmann, M., Kalso, E., & Hari, R. (2007). The compassionate brain: humans detect intensity of pain from another's face. *Cereb Cortex, 17*(1), 230-237.
Samson, D., Apperly, I. A., Braithwaite, J., Andrews, B., & Scott, S. (2010). Seeing It Their Way: Evidence for Rapid and Involuntary Computation of What Other People See. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1255-1266.
Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nat Neurosci, 7*(5), 499-500.
Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage, 19*(4), 1835-1842.
Saxe, R., & Powell, L. (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), 1391-1399.
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 Dev, 80*(4), 1197-1209.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition, 88*(3), B11-21.

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., & Aharon-Peretz, J. (2003). Characterization of empathy deficits following prefrontal brain damage: the role of the right ventromedial prefrontal cortex. *J Cogn Neurosci, 15*(3), 324-337.

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Ann N Y Acad Sci, 1156*, 81-96.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*(5661), 1157-1162.

Sturmer, B., Siggelkow, S., Dengler, R., & Leuthold, H. (2000). Response priming in the Simon paradigm. A transcranial magnetic stimulation study. *Exp Brain Res, 135*(3), 353-359.

Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporo-parietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgment. *Proc Natl Acad Sci U S A, 107*, 6753-6758.

Young, L., Dodell-Feder, D., & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia, 48*(9), 2658-2664.

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40*, 1912-1920.

Young, L., & Saxe, R. (2009). Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47*(10), 2065-2072.

Zwickel, J., Grosjean, M., & Prinz, W. On interference effects in concurrent perception and action. *Psychol Res, 74*(2), 152-171.

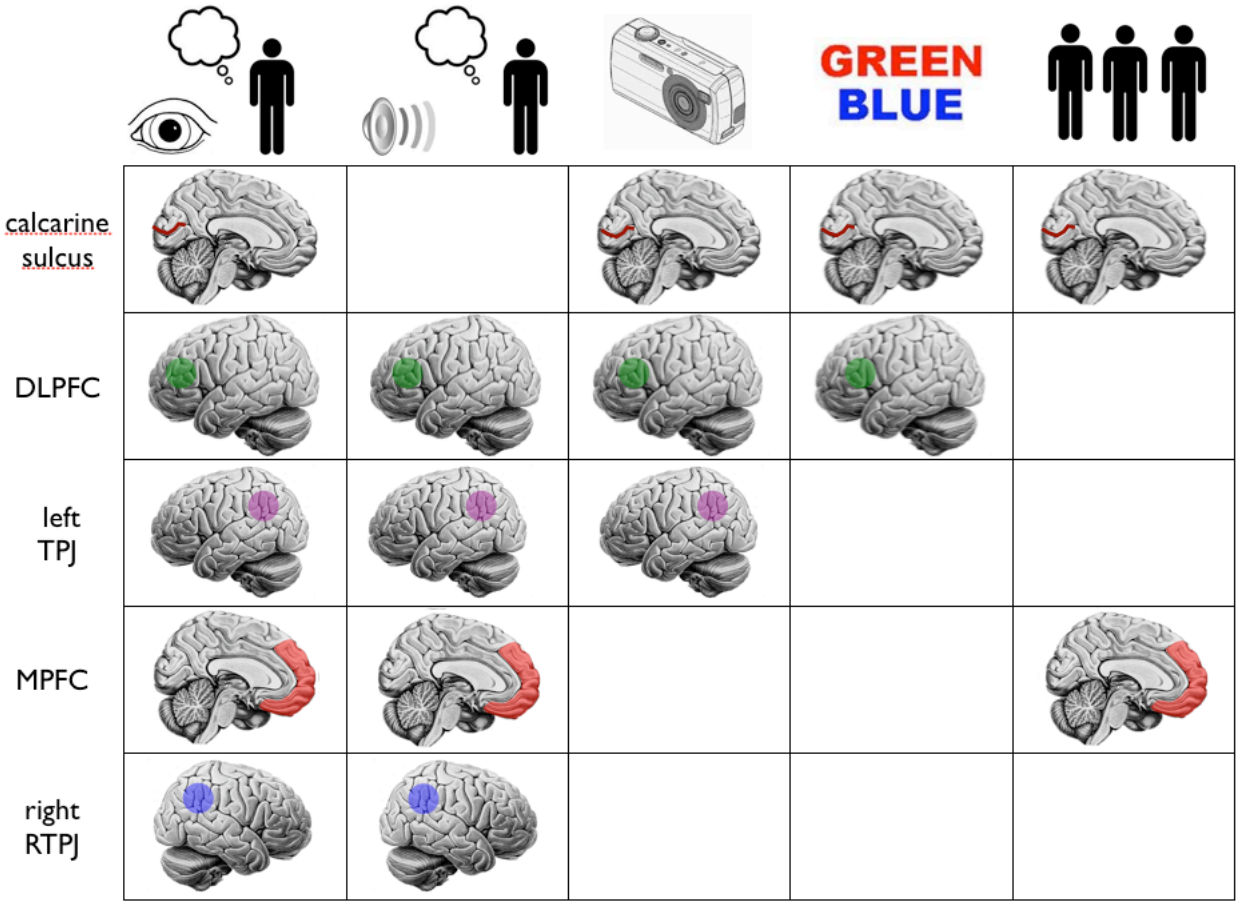
Zwickel, J., Grosjean, M., & Prinz, W. (2010a). On interference effects in concurrent perception and action. *Psychol Res, 74*(2), 152-171.

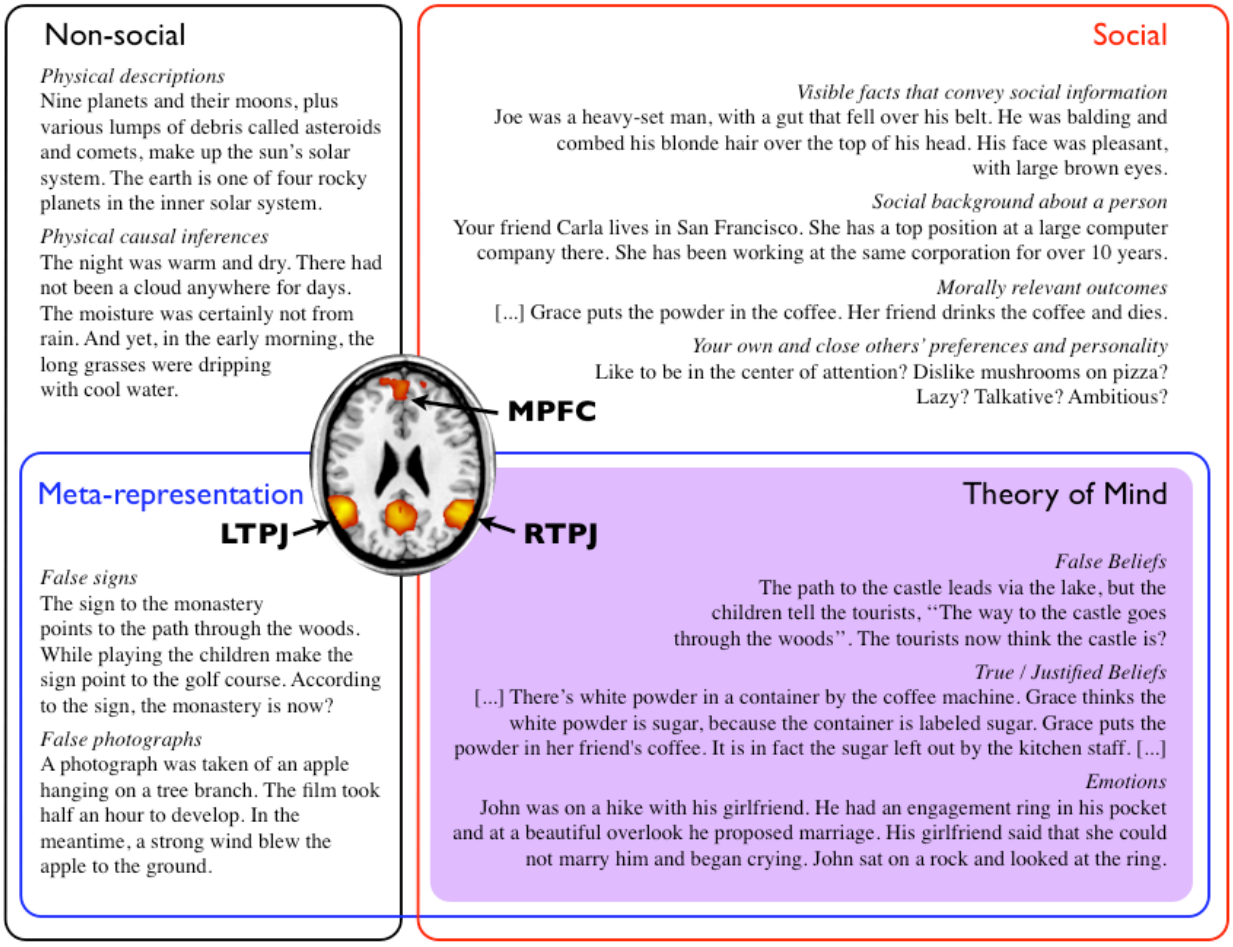
Zwickel, J., Grosjean, M., & Prinz, W. (2010b). What part of an action interferes with ongoing perception? *Acta Psychologica, 134*, 403-409.

Figure Legends

1. Many different brain regions are involved when people perform "Theory of Mind" tasks, for different reasons. The differences between brain regions would be revealed by an imaginary meta-experiment. For example, five different brain regions (rows) would reveal different patterns of functional response across five categories (columns, left to right): (1) visually presented stories depicting people's thoughts, (2) the same stories presented aurally, (3) non-mental meta-representations (e.g., stories about photographs, maps, signs), (4) a Stroop task manipulation of cognitive conflict, and (5) socially-relevant (but non-mental) information about people. These distinct "functional profiles" of response could then be used to infer the function of each of these regions.

2. Sample stimuli from experiments that revealed the functional profile of three brain regions involved in Theory of Mind. The left temporo-parietal junction (LTPJ) shows a higher response when reading stories that require thinking about representations, whether mental (like thoughts) or physical (like signs), compared to stories with no such meta-representational demands. Regions in the medial prefrontal cortex (MPFC) show a higher response when the stories contain socially-relevant information about people. The right temporo-parietal junction is more selective than either, responding when the story contains descriptions of a range of different thoughts, beliefs, desires, or emotions, but not otherwise. Sample stimuli from (Saxe and Kanwisher 2003, Saxe and Wexler 2005, Saxe and Powell 2006, Moran et al 2006, Perner et al 2009, Young et al 2010, Bruneau et al submitted).


GREEN
BLUE
calcarine sulcus
DLPFC
left TPJ
MPFC
right RTPJ


Non-social
Physical descriptions
Nine planets and their moons, plus various lumps of debris called asteroids and comets, make up the sun's solar system. The earth is one of four rocky planets in the inner solar system.
Physical causal inferences
The night was warm and dry. There had not been a cloud anywhere for days. The moisture was certainly not from rain. And yet, in the early morning, the long grasses were dripping with cool water.
Social
Visible facts that convey social information
Joe was a heavy-set man, with a gut that fell over his belt. He was balding and combed his blonde hair over the top of his head. His face was pleasant, with large brown eyes.
Social background about a person
Your friend Carla lives in San Francisco. She has a top position at a large computer company there. She has been working at the same corporation for over 10 years.
Morally relevant outcomes
[...] Grace puts the powder in the coffee. Her friend drinks the coffee and dies.
Your own and close others' preferences and personality
Like to be in the center of attention? Dislike mushrooms on pizza? Lazy? Talkative? Ambitious?
MPFC
Meta-representation
LTPJ
RTPJ
Theory of Mind
False signs
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?
False photographs
A photograph was taken of an apple hanging on a tree branch. The film took half an hour to develop. In the meantime, a strong wind blew the apple to the ground.
False Beliefs
The path to the castle leads via the lake, but the children tell the tourists, "The way to the castle goes through the woods". The tourists now think the castle is?
True / Justified Beliefs
[...] There's white powder in a container by the coffee machine. Grace thinks the white powder is sugar, because the container is labeled sugar. Grace puts the powder in her friend's coffee. It is in fact the sugar left out by the kitchen staff. [...]
Emotions
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.