Philos Stud (2009) 144:447–456
DOI 10.1007/s11098-009-9353-2

The neural evidence for simulation is weaker than I think you think it is

Rebecca Saxe

Published online: 22 April 2009

Abstract Simulation theory accounts of mind-reading propose that the observer generates a mental state that matches the state of the target and then uses this state as the basis for an attribution of a similar state to the target. The key proposal is thus that mechanisms that are primarily used online, when a person experiences a kind of mental state, are then co-opted to run Simulations of similar states in another person. Here I consider the neuroscientific evidence for this view. I argue that there is substantial evidence for co-opted mechanisms, leading from one individual's mental state to a matching state in an observer, but there is no evidence that the output of these co-opted mechanisms serve as the basis for mental state attributions. There is also substantial evidence for attribution mechanisms that serve as the basis for mental state attributions, but there is no evidence that these mechanisms receive their input from co-opted mechanisms.

Keywords Theory of mind · Simulation theory · Mirror neurons · Mentalizing · Temporo-parietal junction · Medial prefrontal cortex

A man stops at the gym on his way to work early on Monday morning, but the gym is closed, so he heads to work. Now consider this extra information: when he sees the gym is closed, he remembers that it's a national holiday; earlier, he had forgotten about the holiday (Apperly 2008). The extra information describes mental states—the man's (later) true and (earlier) false beliefs. These features of the scenario can't be perceived directly by external observers. Even for the man himself, the later realization may be conscious, but the earlier forgetting did not have any observable phenomenal experience. There is nothing it feels like to be forgetting that it's a holiday. Nevertheless, mental states like forgetting are a

R. Saxe (✉)
Brain and Cognitive Sciences, MIT, 46-4019 MIT, 43 Vassar St, Cambridge, MA 02139, USA
e-mail: saxe@mit.edu

common topic of consideration and conversation for human observers of other people's actions and of their own (e.g. "I always forget about holiday Mondays!").

How do human beings use unobservable mental states like "forgetting it's a holiday" to explain and predict actions, both of other people and of their own? This is the problem of "mind-reading". *Simulating Minds* is, as its subtitle claims, a book about the philosophy, psychology and neuroscience of mind-reading, and as such, is no mean feat. For future researchers in all three disciplines, this book will help set the standard for the breadth and detail of empirical results that must be accommodated by any true theory of mind-reading.

Goldman's main goal here is to present the case for Simulation Theory (ST) as an account of mind-reading. The central idea of ST is that for mind-reading, the observer uses her own mind as an analogue model of the mind of target. More specifically, as Goldman presents it here, mind-reading depends on generating in the observer a mental state that (at least partly) matches the state of the target; the observer can then categorise her own state, and use this as the basis for an attribution of a similar state to the target.

According to ST, an act of mind-reading depends on two kinds of mechanisms. The first group are the mechanisms that are "primarily" used online, when a person experiences a kind of mental state; these mechanisms are then co-opted to run Simulations of similar states in another person. I will call these "co-opted mechanisms". The second group are the mechanisms that transform the outputs of Simulation (or any other process) into a mental state attribution to the observer. I will call these "attribution mechanisms."

The distinctive component of ST, as an account of mind-reading, is the reliance on co-opted mechanisms, and so Goldman's discussion, like most experimental tests of ST, have focused on evidence for such mechanisms. This aspect of ST predicts activation of neural mechanisms for experiencing a class of mental states, in the service of attributing a similar state to another person. For neuroscientific evidence to support ST, as Goldman recognizes, it is not sufficient to show that there is "a systematic, repeatable causal pathway, leading from one individual's mental state to a matching (or semi-matching) state in an observer." In addition to co-opting a mechanism for first-person experience, the observer must also use her own mental state as the basis for imputing an instance of this mental state category to the target. So evidence for ST must meet two criteria: (1) there must be evidence for a co-opted mechanism, and (2) the outputs of this mechanism must serve as the basis for a mental state attribution.

My main goal in these comments will be to consider the evidence for this view from one of the methodologies: neuroscience. Neuroscience is relatively new to the mind-reading debate, but the influence of neuroscientific methods is on the rise. I will argue that the neuroscientific evidence, in spite of all the hype, does not support ST.

The general line of argument in what follows will be that Goldman's two criteria have never been met. In the first section, I will argue that there is substantial evidence for co-opted mechanisms, leading from one individual's mental state to a matching state in an observer, but there is no evidence that the output of these co-opted mechanisms serve as the basis for mental state attributions. In the second section, I will argue that there is also substantial evidence for attribution mechanisms that serve as the basis for mental state attributions, but there is no evidence that these mechanisms receive their input from co-opted mechanisms.

1 Co-opted mechanisms

The best evidence for co-opted mechanisms come from two examples. There is evidence that observers activate their own motor control systems, when observing other's actions, and their own representations of two basic emotions, fear and disgust, when observing a corresponding facial expression on someone else's face. In both of these cases, the mechanisms appear to fit the criteria for co-option. However, in both cases, the mechanisms fall short of *mind-reading*, because there is no evidence that these mechanisms serve as the basis for attributing mental states to the target.

A large burst of neuroscience research on mind-reading was sparked by the discovery of "mirror neurones" in macaque monkeys, especially following Goldman's influence article co-authored with the neuroscientist Vittorio Gallese (Gallese and Goldman 1998). The premotor cortex of macaque monkeys contains neurones that code whole action sequences (by contrast to primary motor cortex neurones, which code simpler motor primitives). For instance, neurones in areas F5 and F6 fire when the monkey reaches to grasp an object (Rizzolatti et al. 1990). The critical discovery is that many of these 'motor' neurones also have visual response properties. So-called "mirror-neurones" fire equally when the monkey executes a particular action, and when the monkey observes someone else executing the same action (di Pellegrino et al. 1992; Gallese et al. 1996). There is good evidence for similar mechanisms in human brains (Strafella and Paus 2000; Gangitano et al. 2001).

There is also behavioural evidence that the mechanisms for action planning and action observation are causally inter-twined. When an observed action is incongruent with the action that must be simultaneously executed, execution is seriously impaired. The interference produced by incompatible gestures is much larger than that produced by other kinds of distractors (Brass et al. 2001), even when the hand gestures are task-irrelevant and the other distractors are potentially task-relevant (Sturmer et al. 2000). This interference suggests that the representation of an observed action is competing for the same resources that are necessary for executing one's own action.

Pre-motor mirror neurons thus possess critical features of a co-opted mechanism for ST: they are used during both action execution and observation, and they are (plausibly) primarily "for" the actor's own action planning. At a minimum, mirror neurons are causally involved in the animal's on-line action planning; stimulating a mirror neuron elicits a coherent action sequence (Rizzolatti et al. 1990). More importantly, given their anatomical position and cross-species homologues, the neurones in pre-motor cortex plausibly originally evolved to satisfy the organism's own motor needs. Their use in mind-reading on this view would be a secondary re-deployment of existing resources, as predicted by ST.

In spite of all this evidence, mirror neurones face a critical challenge to a candidate mechanism for mind-reading: there is no evidence that mirror neurons represent the internal states of the target (e.g. his intention) rather than the external properties of the action. The possibility of a role for mirror neurones in mind-reading was initially raised by Gallese and Goldman (1998). Gallese has since claimed that mirror neurons are the mechanism by which "we assign goals, intentions, or beliefs to the inhabitants of our social world" (Gallese et al. 2004). Goldman, by contrast, does not endorse this strong view here. In his discussion of evidence that mirroring could serve as mechanism for mind-reading, he does not discuss mirror neurones at all. Other authors do, though, so it's worthwhile to look briefly at the evidence.

Two experiments in particular are frequently taken as evidence that mirror neurones represent internal states of the target, rather than external features of the action. In one, a monkey watched an experimenter execute a predictable sequence of actions: first, reach and grasp a piece of food, and second, move the food either to a container ("reaching to place") or to his mouth ("reaching to eat"). Some mirror neurones neurons differentiated between the two actions even before they were visibly different (Fogassi et al. 2005). The second experiment (Umilta et al. 2001) used as experimental leverage an empirical fact about monkey mirror neurons: they respond to object-directed actions (e.g. grasping an object) but not to mimes of the same action, in the absence of a target object. The monkey watched object-directed and mimed grasping actions in which the final phase of the action was occluded by the screen. At the end of the sequence, the object-directed and mimed actions were thus visually identical: the person's arm reached down behind the screen. Nevertheless, mirror neurons fired in response to the occluded object-grasps, and not to occluded mimes.

These experiments provide elegant evidence that mirror neurons represent relatively abstract properties of observed actions, going beyond the currently visible movements. The question is: are these abstract properties the internal states of the target? In both experiments, the properties that distinguished the two actions were actual physical features of the action sequence in the environment. Mirror neurones may contain representations of action sequences that make fine-grained predictions about an unfolding action (Csibra 2007), but only in terms of the physical movement, not the internal states (or, a fortiori, the propositional attitudes).

These competing interpretations could be tested. Imagine an alternative version of Umilta et al's paradigm: The actor looks, and sees an apple on the table, and then her view is blocked. While the monkey but not the actor can see, the apple is removed. Immediately, the actor reaches towards the table (her view of the apple's location is still blocked). The physical sequence of the action is the same as in the mime sequence: a reach towards an empty table. For a human observer, though, the actor's intention is obviously the same as if the apple was present: to grasp the apple. Would mirror neurones respond to this action? I predict they would not.

Mirror neurons thus fall short of evidence for ST because there is no evidence that they serve as the basis for attributing any internal state to the target. Instead, they may represent only the external sequence of actions. A similar limitation applies to the other well-studies cases of co-opted mechanisms: (1) fear experience and perception in the amygdala, and (2) disgust experience and perception in the insula.

Patients with amygdala damage have specific difficulties in recognising fear in facial emotional expressions (Adolphs et al. 1994, 1999); these same patients show reduced physiological symptoms of fear in laboratory contexts (Davis and Whalen 2001), and increased risk taking in their everyday lives. Similarly, lesions in the insula cause reduced experience of disgust, and impaired recognition of disgusted facial expressions (Calder et al. 2000). The amygdala and the insula thus seems like excellent candidates for "co-opted mechanisms": they have primary first-person functions in the detection of threatening or aversive stimuli, and they are additionally used during recognition of similar experiences in others.

Again, though, the evidence that attributions of fear or disgust depend on these shared mechanisms is weak. Amygdala damage does not cause impairments in recognising fear from static or dynamic body postures (Atkinson et al. 2007) or from social context (Adolphs and Tranel 2003). Nor do these patients lack all knowledge about facial expressions of fear; they can produce normal facial expressions of fear based on verbal instructions (Anderson et al. 2000). As Goldman notes, even the deficit in facial fear recognition may be completely explained by the patients' tendency not to look at the eye region of the face. Explicitly instructing SM (a woman with bilateral amygdala damage) to look at the eye-region completely removed her deficit. Finally, amygdala damage does not impair inferences about the kinds of situations that cause fear (Adolphs et al. 1995).

In all, the amygdala and insula appear to subserve "sharing" of certain basic emotional experiences. However, the outputs of these regions do not form the basis of mind-reading of emotional mental states (Zaki et al. 2008). By contrast, there is extensive evidence concerning the mechanisms that *are* used for mental state attribution. In these cases, though, there is no evidence that the attributions are based on input from co-opted mechanisms.

2 Attribution mechanisms

To identify neural mechanisms selectively necessary for mental state (in this case, propositional attitude) attribution, Samson and colleagues developed a series of elegant non-verbal false belief tasks. In one set of tasks (reality known), an object was moved without a character's knowledge, but the participants themselves always knew the true location of the object. Passing these tasks required both the ability to represent the character's belief, and the ability to inhibit the participants' own knowledge. In a second group of tasks (reality unknown), an object was moved without the character's knowledge, but the participant also did not know the true location of the object.

Goldman describes one of the studies in this sequence. WBA, a patient with left lateral frontal damage, was selectively impaired on reality-known (i.e. high inhibition) false belief tasks (Samson et al. 2005). That is, when WBA himself believed or desired something, he had trouble inhibiting this belief or desire, in order to make a competing attribution to another person. By contrast, when he did not have a relevant competing belief or desire (as in the reality-unknown false belief tasks), WBA had no trouble correctly attributing a false belief to another person. Also, WBA's deficits were not limited to mind-reading. He failed a whole range of inhibitory control tasks. These results provide compelling evidence that successful mental state attribution frequently depends on the capacity to inhibit one's own competing beliefs and desires.

However, Goldman does not discuss a critical contrast population. A group of patients with left temporo-parietal junction (TPJ) damage failed both reality-known and reality-unknown false belief tasks (Apperly et al. 2004; Samson et al. 2004). These participants passed all of the memory and inhibition control trials of the false belief tasks, and did not fail general tests of inhibitory control outside of the domain of ToM. Samson and colleagues suggested that damage to the left TPJ impairs a component of the concept of mental states, and/or the capacity for meta-representation, independent of the inhibition of one's own mental states.

To test this hypothesis, Samson and colleagues created a more subtle version of the reality-known false belief tasks (Samson et al. 2007). They proposed that left frontal lesions (as in WBA) and left TPJ lesions (patient PF) would cause errors on reality-known false belief tasks for different reasons, leading to systematically different *kinds* of errors. In an example scenario, a neighbour (N) watched through a window as a person (P) hid one object (e.g. a passport) inside a recognizable container (e.g. a pizza box). Then N left, and P switched the object inside the box to a new object (e.g. scissors). N then returned and peered through the window at the box. Participants were asked: what does N think is inside the box?

On this trial, the correct answer would be "a passport" (N's false belief). Note, though, that here are two possible wrong answers: the actual contents of the box (reality error—the scissors), or the probable content of the box based on N's current perspective (appearance error—pizza). As predicted, the two patients made systematically different kinds of errors on this task. WBA made only reality errors, and he made these errors on almost all of the trials. The left TPJ patient, PF, made mostly appearance errors.

In all, Samson and colleagues results reveal two different mechanisms that contribute to propositional attitude attribution, but neither one of them has the characteristics of a co-opted mechanism. The left frontal region appears to support domain-general inhibitory control. The left TPJ region appears to be necessary for formulating meta-representations. Neither of these regions is necessary for the participant to form or use their own propositional attitudes (e.g. beliefs, knowledge, false beliefs) about, for example, the contents of the box. All of these patients were at ceiling in reporting the box's true contents. With respect to these conclusions, the lesion results converge perfectly with a second group of studies using functional neuro-imaging.

One of the most striking discoveries of recent human cognitive neuroscience is that there is a group of brain regions in human cortex that are selectively and specifically recruited during high-level mind-reading. Most studies require participants to attribute false beliefs to people in stories or cartoons (Fletcher et al. 1995; Gallagher et al. 2000; Saxe and Kanwisher 2003; Perner et al. 2006; Gobbini et al. 2007). Four cortical regions are recruited during the false belief condition of each study, relative to controls: right and left temporo-parietal junction (TPJ), medial parietal cortex (including posterior cingulate and precuneus, PC), and medial prefrontal cortex (MPFC).

Of these regions, the right temporo-parietal junction (RTPJ) in particular appears to be selective for mental state attribution. For example, its response is high when subjects read stories that describe a character's thoughts and beliefs but low during stories containing other socially relevant information (e.g. a character's physical appearance, cultural background, or even internal subjective sensations such as hunger or fatigue; (Saxe and Wexler 2005; Saxe and Powell 2006).

Consistent with Samson's lesion studies, neuro-imaging results suggest that false belief task performance depends on distinct domain-general inhibitory control mechanisms, and domain specific mechanisms for mental state attribution. For example, Saxe, Schulz and Jiang (Saxe et al. 2006) compared responses during stories about beliefs versus physical representations (e.g. photographs and maps) to identify brain regions implicated in belief reasoning, and a difficult versus easy response-selection task to identify brain regions implicated in domain-general response selection and inhibitory control. Reasoning about beliefs provoked robust activity in the brain regions associated with inhibitory control (including intra-parietal sulcus and frontal eye-fields)—but an equal response was observed in those regions during photograph stories. That is, domain-general mechanism for inhibitory control, response selection, and so on are recruited by both belief and photograph stories to equivalent degrees, and there are additional brain regions that are recruited only for beliefs.

Could any of these regions constitute a co-opted mechanism for propositional attitudes? Some authors propose that they could. For example, Vogeley et al. (2001) required subjects to attribute hypothetical mental states to "themselves", in order to explain hypothetical actions (Vogeley et al. 2001). Participants read short verbal stories about a protagonist. Half of the stories were presented in the second person (e.g. "In the morning, when you leave the hotel, the sky is blue and the sun is shining. So you do not expect it to start raining.") Since these stories described non-actual events and actions, the participants were not experiencing the mental states described in these stories. Instead the participants must infer and attribute the mental states that would cause the described actions. Vogeley et al. (2001) reported that these second-person stories elicited a strong response in the TPJ and associated regions—at least as strong as the response to stories in the third person.

Vogeley and colleagues' results provide lovely evidence that mental state attribution to the self, and to other people, depend on shared mechanisms. Critically, though, this experiment provides no evidence for a co-opted mechanism. The key prediction of ST is that people use the same mechanism for *having/forming* propositional attitudes themselves, and for attributing those attitudes to other people. In Vogeley's experiment, in the "self" condition, the participants did not actually form the belief that it would rain. They attributed to themselves the hypothetical belief that it would not rain.

Similarly, a whole series of studies have shown that a region in MPFC is recruited for attributing personality traits to the self, and to well-known or similar others (Kelley et al. 2002; Macrae et al. 2004; Mitchell et al. 2005; Mitchell et al.

2006; Jenkins et al. 2008). Again, these results suggests that concepts of self and familiar others have a shared neural basis. However, these results do not provide evidence for ST [contrary to claims by the authors, e.g. (Mitchell et al. 2005; Decety and Grezes 2006)]. A co-opted mechanism for personality attribution would have to be primarily used for *being* lazy or courageous or friendly, and secondarily for attributing laziness, courage or friendliness to others. Instead, all of the evidence suggests that the MPFC is recruited specifically when participants self-attribute laziness, not when they are being lazy.

More generally, many claims that neuro-imaging evidence supports ST have confused two different aspects of one's own mental states (Apperly 2008): the first-person, on-line possession of a mental state, versus the attribution of mental states to the self (in the past or hypothetically). The right experiment is possible. Pairs of trials would be designed so that on one trial the participant is induced to acquire a specific propositional attitude, and then, on the other trial, to attribute the same propositional attitude to another person. This experiment has not yet been done.

In all, there's no evidence yet for a co-opted mechanism for propositional attitude attribution. On the contrary, all the hints suggest that there are distinct mechanisms for propositional attitude attribution, and they are not co-opted.

Unlike many neuroscientists, Goldman is carefully aware of the distinction between the first-person experience of a mental state and the self-attribution of the same mental state. I agree with him that a clear and plausible account of self-attribution is a critical test of any theory of mind-reading. On the one hand, we clearly know our own minds differently than we know other minds, in at least some respects; on the other hand, there are many contexts in which self-attribution appears to share both the mechanisms and the flaws of mind-reading for other people (Saxe 2005). As Goldman shows, neither ST nor any of its competitors can yet provide a satisfactory resolution of this tension.

As for the contribution from neuroscience, I agree with Goldman that "neuroscientists have established the existence of a wide range of inter-personal mirroring mechanisms, in which the cognitive states of one organism are matched or mirrored by similar cognitive states in an observing organism." But so far none of these matching mechanisms fit the criteria for a distinctively Simulation-based mechanism of mind-reading.

When the man gets to the gym and sees that it is closed, there are some neural mechanisms in him that allows him to see the gym, recognise that it's closed, and remember, based on this information, that it's a holiday Monday. At that same moment, the man may also attribute to himself the past false belief that it isn't a holiday. The neural evidence suggests that such self-attributions depend on a completely distinct set of brian regions than the first-person formation of the corresponding mental states: the TPJ, among others. An observer across the street, seeing a guy standing with his gym bags in front of the darkened building, could make the same attribution: "He must have forgot it was a holiday Monday!" If so, the observer, like the man himself, would recruit the TPJ during this attribution. But the TPJ is not a co-opted forgetting mechanism; it is a domain-specific mental state attribution mechanism. That is, in older terminology, the TPJ is a neural mechanism of a Theory of Mind.

Acknowledgements The author is grateful to Liane Young, Hyo Won Gweon, Laura Schulz, Marina Bedny, Jonathan Scholz and Allan Adams for helpful conversations and moral support during the preparation of this manuscript.

References

Adolphs, R., & Tranel, D. (2003). Amygdala damage impairs emotion recognition from scenes only when they contain facial expressions. *Neuropsychologia, 41*(10), 1281–1289.

Adolphs, R., Tranel, D., et al. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature, 372*(6507), 669–672.

Adolphs, R., Tranel, D., et al. (1995). Fear and the human amygdala. *Journal of Neuroscience, 15*(9), 5879–5891.

Adolphs, R., Tranel, D., et al. (1999). Recognition of facial emotion in nine individuals with bilateral amygdala damage. *Neuropsychologia, 37*(10), 1111–1117.

Anderson, A. K., Spencer, D. D., et al. (2000). Contribution of the anteromedial temporal lobes to the evaluation of facial emotion. *Neuropsychology, 14*(4), 526–536.

Apperly, I. A. (2008). Beyond Simulation-Theory and Theory-Theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind". *Cognition, 107*(1), 266–283.

Apperly, I. A., Samson, D., et al. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience, 16*(10), 1773–1784.

Atkinson, A. P., Heberlein, A. S., et al. (2007). Spared ability to recognise fear from static and moving whole-body cues following bilateral amygdala damage. *Neuropsychologia, 45*(12), 2772–2782.

Brass, M., Bekkering, H., et al. (2001). Movement observation affects movement execution in a simple response task. *Acta Psychologica, 106*, 3–22.

Calder, A. J., Keane, J., et al. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience, 3*(11), 1077–1078.

Csibra, G. (2007). Action mirroring and action interpretation: An alternative account. In P. Haggard, Y. Rosetti, & M. Kawato (Eds.), *Sensorimotor foundations of higher cognition. Atterntion and performance XXII* (pp. 435–459). Oxford: Oxford University Press.

Davis, M., & Whalen, P. J. (2001). The amygdala: Vigilance and emotion. *Molecular Psychiatry, 6*(1), 13–34.

Decety, J., & Grezes, J. (2006). The power of simulation: Imagining one's own and other's behavior. *Brain Research, 1079*(1), 4–14.

di Pellegrino, G., Fadiga, L., et al. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91*(1), 176–180.

Fletcher, P. C., Happe, F., et al. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*(2), 109–128.

Fogassi, L., Ferrari, P. F., et al. (2005). Parietal lobe: From action organization to intention understanding. *Science, 308*(5722), 662–667.

Gallagher, H. L., Happe, F., et al. (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.

Gallese, V., Fadiga, L., et al. (1996). Action recognition in the premotor cortex. *Brain, 119*(Pt 2), 593–609.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493–501.

Gallese, V., Keysers, C., et al. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8*(9), 396–403.

Gangitano, M., Mottaghy, F. M., et al. (2001). Phase-specific modulation of cortical motor output during movement observation. *Neuroreport, 12*(7), 1489–1492.

Gobbini, M. I., Koralek, A. C., et al. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience, 19*(11), 1803–1814.

Jenkins, A. C., Macrae, C. N., et al. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences of the United States of America, 105*(11), 4507–4512.

Kelley, W. M., Macrae, C. N., et al. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience, 14*(5), 785–794.

Macrae, C. N., Moran, J. M., et al. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex, 14*(6), 647–654.

Mitchell, J. P., Banaji, M. R., et al. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience, 17*(8), 1306–1315.

Mitchell, J. P., Macrae, C. N., et al. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*(4), 655–663.

Perner, J., Aichorn, M., et al. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience, 1*(3–4), 245–258.

Rizzolatti, G., Gentilucci, M., et al. (1990). Neurons related to reaching-grasping arm movements in the rostral part of area 6 (area 6a beta). *Experimental Brain Research, 82*(2), 337–350.

Samson, D., Apperly, I. A., et al. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nature Neuroscience, 7*(5), 499–500.

Samson, D., Apperly, I. A., et al. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain, 128*(Pt 5), 1102–1111.

Samson, D., Apperly, I. A., et al. (2007). Error analyses reveal contrasting deficits in "theory of mind": Neuropsychological evidence from a 3-option false belief task. *Neuropsychologia, 45*(11), 2561–2569.

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences, 9*(4), 174–179.

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. 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., Schulz, L., et al. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Social Neuroscience, 1*(3–4), 284–298.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43*(10), 1391–1399.

Strafella, A. P., & Paus, T. (2000). Modulation of cortical excitability during action observation: A transcranial magnetic stimulation study. *Neuroreport, 11*(10), 2289–2292.

Sturmer, B., Siggelkow, S., et al. (2000). Response priming in the Simon paradigm. A transcranial magnetic stimulation study. *Experimental Brain Research, 135*(3), 353–359.

Umilta, M. A., Kohler, E., et al. (2001). I know what you are doing. A neurophysiological study. *Neuron, 31*(1), 155–165.

Vogeley, K., Bussfeld, P., et al. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage, 14*(1 Pt 1), 170–181.

Zaki, J., Bolger, N., et al. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19*(4), 399–404.