PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2009) **364**, 1309–1316
doi:10.1098/rstb.2008.0318

Inferences about mental states

Jason P. Mitchell*

Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02138, USA

Human social cognition relies on an ability to predict what others will think, feel or do in novel situations. Research in social neuroscience has consistently observed several brain regions that contribute ubiquitously to these abilities, including medial prefrontal cortex and aspects of lateral and medial parietal cortex. Interestingly, parallel work has suggested that this same network of regions subserves several seemingly distinct phenomena—notably, the abilities to remember the past, imagine the future and visualize spatial layouts—suggesting the existence of a common set of cognitive processes devoted to projecting oneself into worlds that differ mentally, temporally or physically from one's current experience. This use of self-projection to understand others' minds requires perceivers to solve three distinct cognitive challenges: (i) generating a simulated facsimile of one's own hypothetical mental states in a given situation, (ii) suppressing one's own current mental states, and (iii) deciding on the appropriateness of simulated states for understanding a particular other person. The present paper reviews recent psychology and neuroscience research aimed at understanding the underlying mechanisms that allow humans to solve each of these cognitive challenges to use self-projection to predict and understand the mental states of others.

Keywords: social neuroscience; self; mentalizing

1. INTRODUCTION

What does it take to read the mind of another person? Although the idea of mind-reading implies the possession of supernatural abilities or exotic technology, even ordinary, humdrum humans are artfully accomplished telepaths. We routinely rummage around in the minds of others, deciphering what those around us are feeling and thinking, inferring what others intend and desire, and constructing an understanding of others' stable dispositions and personality traits. In addition, just like the most powerful telepaths, we not only passively read, but also actively control other minds, influencing what others think, feel and do through our communicative acts and social behaviour.

So how exactly do humans gain access to the inner workings of others' minds? After all, mental states remain safely sequestered inside one's own head, invisible to direct inspection by others. None of us has ever directly caught sight of another person's thoughts, feelings or intentions but must instead infer others' mental states obliquely, from sources other than direct perception. Uncovering the processes of indirect inference, deduction and guesswork through which perceivers understand the minds of others is one of the central pursuits in the study of social cognition.

In recent years, commentators have suggested that either of two broad classes of cognitive process could serve to generate suitable inferences about other minds. The first class—known generically as 'theory–theory'—suggests that perceivers may bring to bear a sophisticated set of rules for deciphering the internal workings of other minds. By analogy to language, individuals may learn a finite set of 'primitives' about the way other people work and a 'social grammar' for combining them. For example, most of us have learned that, if possible, people will seek relief from aversive bodily states such as hunger and that they also believe that refrigerators usually contain food. Upon witnessing another person hurry to the refrigerator and wrench open its door, we assume that she may be hungry, rather than, for instance, assuming that he is trying to establish whether the light stays on when the door is shut. Indeed, some theorists have suggested that a fully elaborated body of such social knowledge may develop through an explicit process of generating and testing hypotheses about the factors that guide the behaviour of others, a view of children as 'young scientists' (Gopnik & Wellman 1992, 1994; Wellman 1992; Saxe 2005).

A second class of explanation for human social abilities focuses on the use of one's own knowledge of self as the basis for understanding others. Instead of a rule-based system that outputs inferences about other people, proponents of such 'simulation' or 'projection' accounts suggest that perceivers can use their own mental states as proxies for other minds. Watching someone lunge at the refrigerator, a perceiver might (unconsciously) imagine himself engaging in the same act and try to answer the question, 'what would lead me to engage in that behaviour?'. To the extent that a perceiver more easily imagines rushing to a refrigerator to satiate his hunger than to quell his curiosity about its internal workings, he may presume that the other person is likewise motivated by feeling hungry. Such simulations work even in the absence of information about another person's behaviour; for example, to

*mitchell@wjh.harvard.edu

One contribution of 18 to a Theme Issue 'Predictions in the brain: using our past to prepare for the future'.

predict what another person might think or feel in a hypothetical situation (a person's beloved childhood pet dies on the same day as her wedding), we might imagine experiencing the same constellation of events, predict what we ourselves would subsequently think and feel, and infer that another person would experience roughly those same states (Heal 1986; Goldman 1992; Davies & Stone 1995; Suddendorf & Corballis 2007).

These two strategies of rule-based or self-referential mentalizing have often been conceptualized as mutually exclusive possibilities for how the human mind makes inferences about the mental states of others (Saxe 2005). However, even cursory treatment of other cognitive systems will suggest the implausibility of such an 'either–or' assumption about social cognition. Just as evolution has endowed us with multiple systems for generating a perceptual representation of the physical entities in our environment (vision, audition, olfaction, etc.), humans may well be able to deploy a variety of strategies for generating a representation of the social entities in our environment (Mitchell 2005). Humans (and other animals) flexibly use a number of distinct perceptual inputs depending on both the information available in the physical environment (e.g. in the dark, we rely relatively less on vision) and our specific goals at a given time (e.g. olfaction is particularly useful for distinguishing rotten from edible foods, whereas audition is not). Likewise, it seems likely that, for the purpose of thinking about other minds, humans flexibly employ a number of distinct mentalizing strategies depending on both available information (e.g. whether one can see a target's face) and one's specific social goal (e.g. explaining someone's unexpected behaviour, predicting what someone may do in the future, or choosing a gift that someone will enjoy). Rather than debating which singular process gives rise to human social abilities, a central aim of social cognition should be identification of the full range of available mentalizing processes and a delineation of the contexts in which one or another is brought to bear on the problem of understanding others.

2. THE FUNCTIONAL NEUROANATOMY OF MENTALIZING

In recent years, researchers have advanced this goal through investigating the neural basis of social cognition. By identifying the brain regions engaged by mentalizing, researchers have been able to link the cognitive processes subserving social cognition to those involved in other mental operations. Such studies have established a small and highly reliable network of regions that is preferentially engaged when perceivers mentalize about the minds of others. Most notably, these areas include the medial prefrontal cortex (MPFC), the temporo-parietal junction (TPJ) and the medial parietal cortex. In fact, the observation that consideration of others' mental states produces greater activation in these regions (especially in the MPFC) is one of the most reliable observations in cognitive neuroscience. Researchers have observed this pattern of activation regardless of whether mental state inferences are prompted by stories (Fletcher *et al.* 1995; Saxe & Kanwisher 2003; Saxe & Wexler 2005; Saxe & Powell 2006; Mitchell 2008), cartoons (Castelli *et al.* 2000, 2002; Gallagher *et al.* 2000; Martin & Weisberg 2003; Wheatley *et al.* 2007), in the context of competitive and economic games (McCabe *et al.* 2001; Gallagher *et al.* 2002), or by task instructions to think about a specific person's mind (Goel *et al.* 1995; Mitchell *et al.* 2004, 2005*b*, 2006*a*).

Parallel observations have also been made outside of studies employing neuroimaging methods. Impairments in social knowledge follow lesions to either the MPFC (Stone *et al.* 1998; Bird *et al.* 2004; Shamay-Tsoory *et al.* 2004, 2006) or the TPJ (Apperly *et al.* 2004, 2006). Likewise, mentalizing difficulties arise reliably when transcranial magnetic stimulation is used to create transient 'functional lesions' of the MPFC (Lev-Ran *et al.* submitted) or the TPJ (Young *et al.* submitted). Autism, defined centrally as a social-cognitive deficit (APA 1994; Baron-Cohen 1995), has been linked to decreased metabolic activity in the MPFC (Kennedy *et al.* 2006). Moreover, patients with the frontal variant of frontotemporal dementia, who typically experience gross changes in social behaviour and mentalizing ability, have been shown to have particularly pronounced atrophy in medial frontal regions (Bozeat *et al.* 2000; McKhann *et al.* 2001; Gregory *et al.* 2002; Salmon *et al.* 2003).

Intriguingly, the same pattern of medial frontal, temporo-parietal and medial parietal activity consistently accompanies a number of disparate tasks that, at first blush, appear to share little in common with mentalizing. Most notably, these regions are engaged by attempts to prospectively imagine the future or to retrospectively remember the past (Addis *et al.* 2007; Buckner & Carroll 2007; Schacter *et al.* 2007; Spreng *et al.* in press). For example, Addis *et al.* asked participants alternately to imagine their future self experiencing a specific event (cued by an object, such as 'dress') or to recall an actual event that occurred in their past. Both prospection and episodic memory engaged a highly overlapping network of regions that included MPFC, bilateral TPJ and medial parietal cortex. In addition, the same network has also been argued to play a role in spatial navigation (Buckner & Carroll 2007; Spreng *et al.* in press).

The fact that prospection, episodic memory, spatial navigation and mentalizing each draws on the same set of brain regions suggests that each likewise draws on a common set of cognitive processes. What cognitive challenge might these four disparate tasks share? One answer to this question is that each requires perceivers to conjure up a world other than the one that they currently inhabit: prospection obliges perceivers to imagine possible future scenarios; episodic memory relies on the reconstruction of bygone events; and spatial navigation often includes simulations of possible routes between locations. In other words, prospection and episodic memory can be conceived of as forms of mental time travel, and spatial navigation as a form of mental teleportation, all of which depend critically on the ability to project oneself outside of the here and now, imagining times or locations other than the one currently being experienced (Suddendorf & Corballis 2007).

What then of inferring others' mental states? That mentalizing also relies on the MPFC/TPJ/medial parietal network suggests that understanding other minds may likewise require mentally projecting oneself into a scenario other than the one currently being experienced. Whereas the other three forms of projection require imagining oneself inhabiting a different temporal or physical location, mentalizing may draw on one's ability to imagine oneself inhabiting a different *mental* situation: the mind of another person. That is, mentalizing may rely heavily on the same kinds of constructive, projection-based processes as do other tasks that draw on this network of brain regions.

In making use of such projective strategies for social cognition, perceivers may construct a mental representation of a target's experience (i.e. one different from that being currently experienced first-hand), predict the kinds of thoughts and feelings they may have in such a situation and then assume that the target of mentalizing will think or feel much the same thing. Provocatively, this recipe for social cognition suggests that, in using self-projection as a basis for mentalizing, perceivers confront a series of three cognitive challenges that must be overcome before they can use self to understand others. First, perceivers must decide whether a particular target individual really would respond to an experience with the same thoughts and feelings that they predict for themselves. Second, perceivers must be capable of using their simulation as a basis for generating their own possible mental states; that is, they must make reasonable predictions of what they would think or feel in a given situation. Third, perceivers must successfully suppress their own current mental states in favour of the imagined thoughts and feelings of someone else. Each of these three simulation challenges has been addressed by recent work in cognitive neuroscience, reviewed in the following.

3. THE EFFECT OF PERCEIVED SIMILARITY ON SELF-PROJECTIVE MENTALIZING

Using one's predictions about self as a proxy for others only works when one can reasonably assume that another person will have similar responses to a situation. If a perceiver views herself as dissimilar from a target individual, her use of self-projective mentalizing may be wholly inappropriate. Using the involvement of the MPFC as one benchmark for whether one has deployed self-projection as a basis for mental state inference, a number of recent studies have suggested that perceivers do indeed restrict projective mentalizing to those targets perceived as similar to self. In an initial study (Mitchell *et al.* 2005*a*), participants were scanned while making two kinds of judgements about target individuals in a series of photographs. During mentalizing judgements, participants were asked to judge how pleased the target person was to have had her or his photograph taken, whereas in non-mentalizing judgements, participants instead judged how symmetrical the person's face was. After scanning, participants considered each target a second time and were asked to indicate how similar they perceived the person to be to themselves. Analyses identified brain areas in which activity correlated with subsequent ratings of similarity, revealing that a ventral MPFC region responded preferentially to photographs of people perceived to be similar to the participant, but only during the mentalizing task. Importantly, this same region of ventral MPFC has been implicated repeatedly in earlier studies of self-referential thought, in which participants were instructed to report on their own, first-person mental states, preferences or personality traits (Johnson *et al.* 2002; Kelley *et al.* 2002; Macrae *et al.* 2004; Schmitz *et al.* 2004; Vogeley *et al.* 2004; Moran *et al.* 2006).

A second study extended these observations by manipulating targets to have similar or dissimilar ideologies as participants (Mitchell *et al.* 2006*b*). While being scanned, participants judged the possible mental states of three different people: a person with liberal political views; a person with conservative political views; and themselves. Specifically, participants reported how likely each target would be to hold each in a series of opinions and preferences (e.g. 'get frustrated sitting in traffic'). After scanning, the extent to which each subject identified with each target was assessed with a version of the implicit association test (Greenwald *et al.* 1998) that measured whether the participant more closely associated self with the liberal or with the conservative target. Critically, reporting one's own preferences or judging those of the *similar* other was associated with activity in a ventral MPFC region that was nearly identical to the one observed for similar others in the first study, providing additional evidence that perceivers rely on the same processes when thinking about their own preferences or those of a like-minded individual.

Finally, a third study used a repetition suppression paradigm to provide stronger evidence that perceivers consider their own thoughts and feelings when mentalizing about similar others (Jenkins *et al.* 2008). Repetition suppression is the observation that activity in the brain region(s) associated with a given process is typically reduced upon repeated engagement of that process (Grill-Spector *et al.* 2006). To the extent that perceivers spontaneously predict their own mental states when considering those of a like-minded individual, we expected to observe repetition suppression in the ventral MPFC for either repeated judgements of self or self-judgements following judgements of a similar (but not dissimilar) other. Participants made a series of paired mentalizing judgements (e.g. 'how much does the person enjoy skiing?'). In a given pair, participants reported their own preference immediately after judging the preference of one of three targets: a similar other; a dissimilar other; or self. Reporting one's own preference immediately following an initial judgement of self was associated with significant reductions of ventral MPFC activity, demonstrating that this region does indeed show repetition suppression for repeated stimuli. Critically, this same pattern was observed for self-reports that followed judgements of a similar other, suggesting that perceivers also spontaneously computed their own mental states in considering those of a like-minded individual. By contrast, when participants judged self after a dissimilar other, significant activation above

baseline was observed in the ventral MPFC, suggesting that participants did not spontaneously consult their own self-based knowledge when judging a person perceived to be dissimilar.

Together, these data suggest that perceivers selectively deploy self-referential strategies for mentalizing only when they can reasonably use their own mental states as a proxy for those of another (i.e. similar) person. By contrast, when an assumption of self–other similarity cannot be made, perceivers will decline to project themselves into the mental experience of the other person. Of course, these data beg a question about how one goes about inferring the mental states of those perceived to be dissimilar from self, that is, when projection may be inappropriate. Describing the cognitive processes deployed when perceivers decline to mentalize in a self-referential manner remains an unsettled challenge for researchers of social cognition.

4. PREDICTIONS ABOUT ONE'S OWN MENTAL STATES

If we mentalize about each other by imagining ourselves experiencing an event 'as another person' and then predicting our own mental states in that situation, how do we make reasonable predictions about our hypothetical thoughts and feelings in the first place? To use ourselves as a proxy for others' mental states, we must not only be able to imagine ourselves as another person but also be able to simulate richly enough to provoke in us a concomitant set of feelings and thoughts. That is, we must first conjure up the actual feeling states that accompany a particular experience or think the thoughts that might arise in a given scenario before we can proceed to extend those simulated feelings and thoughts to another person.

Surprisingly little cognitive work has addressed the question of how humans predict their own mental states, despite reasonable arguments that this skill represents an important line separating human cognition from the mental systems of other primates (Gilbert 2007). Do perceivers simply liken such experiences to similar situations from their past, thereby drawing on episodic memory to generate a prediction of the kinds of mental states they might encounter? Do they apply a set of rules—somewhat similar to those postulated by theory–theorists for mentalizing about others—that output a prediction about what a perceiver herself would experience in a situation? Or do perceivers engage in something richer and more constructive than either of these two cognitive strategies?

Although we almost certainly make occasional use of both memory and some kind of rule-based inferences for predicting our own mental states, we must also possess a system for predicting our mental states in truly novel situations, where we cannot avail ourselves of memory for our past experiences or global rules about people in general. When asked to answer highly unusual questions, such as whether they would rather spend a year alone as an astronaut on Mars or a year in a submarine stationed under the polar ice cap, respondents can generate an answer that feels as if it accurately reflects their preference. Such questions do not leave us dumbfounded, despite most of us having neither lived on Mars nor under the North Pole (nor knowing anyone who has); moreover, if asked again tomorrow, we would probably provide the same answer. How does our cognitive system predict our mental states under such radically novel and unusual situations?

Across a number of studies, our group has attempted to address this question by examining the neural systems that support stable predictions about a particular form of one's future mental states: one's preferences about the kinds of things one would like or dislike (Ames *et al*. 2008; Jenkins *et al*. 2008; Mitchell *et al*. 2006*b*). In these studies, participants considered a series of questions that asked them about their opinions and preferences across a range of topics. Questions were designed such that respondents were unlikely to be able to answer them on the basis of 'precompiled' semantic representations. Intriguingly, this task consistently engaged one of the regions involved in projective simulation of other minds and other times and places: the ventral MPFC. These observations are suggestive that, when asked to calculate their own preferences, respondents may begin by (either consciously or unconsciously) simulating themselves enmeshed in the relevant situation and reading off the kinds of feelings they expect to have about it. However, additional work is needed to provide additional empirical support for this suggestion more fully.

Consistent with the observation that the ventral MPFC subserves predictions about one's own likes and dislikes, neuropsychological patients with damage to this region show considerable instability in their reported preferences. Fellows & Farah (2007) asked patients with damage to the ventral MPFC, patients with damage to the dorsolateral PFC and healthy controls to indicate how much they liked a series of actors. Actors were presented in pairs, and participants were instructed to report which of the two they preferred. Both healthy controls and patients with dorsolateral PFC damage reported highly stable preferences: if a participant preferred Ben Affleck over Matthew Broderick, and Broderick over Tom Cruise, then he almost always also preferred Affleck over Cruise (i.e. $A > B$ and $B > C$, therefore $A > C$). By contrast, patients with ventral MPFC damage showed much more inconsistent preferences; for example, an individual might indicate that he preferred Affleck over Broderick, Broderick over Cruise, but choose Cruise in a head-to-head comparison with Affleck (i.e. $A > B$, $B > C$, but $C > A$). Although they do not directly address the question of how exactly individuals come to an understanding of their own preferences, these data support the view that doing so draws on the ventral MPFC, and may rely on projective simulations of one's potential experience.

Recently, we have extended this suggestion by demonstrating that individual differences in the tendency to engage the ventral MPFC during judgements of future preferences correlate with rational economic decisions (Mitchell *et al*. submitted). Research in behavioural economics has repeatedly demonstrated that individuals make decisions that maximize happiness in the present at the expense of

one's future enjoyment. For example, when given the choice between $10 now and $12 in one week, people have a tendency to give up the larger, later reward in favour of the immediate payment. Likewise, we have a tendency to commit ourselves to future actions that we will regret when the actual time comes to carry them out. We may anticipate it being fun to travel to a conference or write a chapter in some months' time, but, upon actually having to carry through on our obligation, regret having consigned ourselves to spending time that we now see that we could use in other ways.

Such inaccurate predictions about our distant preferences may result from failure to project ourselves appropriately into the future. That is, we may fail to engage in a rich simulation of the concrete details involved in actually engaging in certain future activities (Trope & Liberman 2003). To the extent both that the ventral MPFC subserves these kinds of simulations and that simulating the distant future will be more difficult than thinking about the here and now, we might expect to observe less activation in this area when perceivers report their preferences for the future compared with the present. We tested these predictions by scanning participants while they predicted how much they would enjoy a series of everyday activities (e.g. 'browsing in a bookshop for 30 min') at one of two time horizons: either 'in the next 24 hours' or 'this time next year'. As predicted, significantly less ventral MPFC activity accompanied predictions about the far future, consistent with the possibility that participants failed to project themselves fully into another time and place when considering distant events.

More intriguingly, the extent to which this ventral MPFC activity differentiated between present and future preferences correlated significantly with participants' unwillingness to wait for a larger payment instead of taking a smaller, immediate one. Specifically, participants were offered the chance to receive a $10 gift certificate immediately or to wait one month for a larger amount of money. Although some participants were willing to wait a month for any larger reward (even $11), most would only wait when the later reward was substantially larger than the immediate pay-off. Critically, the minimum amount necessary for which a participant would wait (a measure of how impatient that participant was for the immediate reward) was directly related to the degree to which the ventral MPFC differentiated between judgements in the present and the future, suggesting that impatient participants may have failed to adequately simulate the experience of receiving the larger reward.

5. SUPPRESSING ONE'S OWN MENTAL STATES

Finally, the use of self-projection as a basis for understanding the mental states of others poses a third challenge to perceivers: how does one discriminate between one's actual thoughts and feelings and those that are merely being simulated? Of course, this distinction between self and other is not always maintained, as anyone who has cried or cringed along with a movie character has experienced first-hand. Nevertheless, despite the richness of our projections into the mental shoes of other people, such confusion between one's own and others' mental states is the exception rather than the rule. When engaged in consideration of the internal workings of another person, the human mind must possess some mechanism for keeping track of which mental states 'belong' to whom.

Although relatively little research has examined this question directly, one hypothesis suggests that one of the regions implicated in the self-projection network specifically keeps track of the differences between simulated and personal mental states, namely the TPJ. This region has been observed consistently when perceivers encounter someone whose mental states conflict with their own, suggesting a role for this region in tracking the differences between self and other. Specifically, Saxe and colleagues have extensively documented the contributions of this region to understanding others' beliefs about the world that differs from what perceivers themselves know to be true. In these studies, perceivers read stories in which a protagonist's knowledge about the world becomes outdated over time (e.g. a child hides his cookies in a cupboard, but then his mother moves them to the cookie jar while he is away at school) and are asked to predict the target's action (e.g. where will he look for his cookies upon returning home?). To answer correctly, perceivers must inhibit what they themselves know to be true about the world (the cookies are actually in the jar) in favour of simulating the mental states of the target (he still thinks that they are safely stashed in the cupboard). Compared with reading logically similar stories about outdated physical representations (such as photographs), suppressing one's own mental states in favour of another's reliably activates the TPJ (Saxe & Kanwisher 2003; Saxe & Wexler 2005; Saxe & Powell 2006; Mitchell 2008), and lesions to TPJ impair reasoning about false beliefs (Samson *et al.* 2004; Apperly *et al.* 2005, 2006).

Interestingly, the same region of the TPJ implicated in reasoning about false beliefs has been observed repeatedly during non-social tasks that require participants to reorient spatial attention away from a salient distractor (Corbetta *et al.* 2000, 2005; Corbetta & Shulman 2002; Shulman *et al.* 2002; Astafiev *et al.* 2003, 2006; Kincade *et al.* 2005; Serences *et al.* 2005). For example, when observers first see a cue that indicates a target stimulus will appear to the left, they covertly shift their attention in that direction; if the stimulus then surprisingly appears to the right, perceivers must inhibit the original location of their attention and reorient it to the actual target location. Indeed, when the same participants alternately performed both this kind of attentional reorienting task and the false belief task, the very same region of the TPJ was engaged (Mitchell 2008), suggesting that these two seemingly disparate tasks actually draw on the same cognitive process. Although the exact identity of this cognitive process has yet to be fully uncovered, it seems likely that it involves a suppression of salient stimuli in favour of a less immediate alternative. To the extent that one's own mental states (e.g. what one knows to be an actual fact) is this kind of highly salient representation, the TPJ may be required to inhibit attention away from one's own beliefs and

reorient them towards those of others. This kind of process may be particularly important in those situations in which one's own mental states are incompatible with a simulated representation of the mind of another person.

6. CONCLUSIONS

Humans possess a unique ability to traffic in the internal mental states of other people. We can infer complicated emotional states in others, understand that others can believe things that are demonstrably false, and parse others' behaviour as a clue to what they are thinking and feeling. Recently, the use of data from neuroimaging and neuropsychological patients has dramatically enhanced our understanding of how humans accomplish these feats of mind-reading. Specifically, a decade of neuroscience research has implicated several regions—the MPFC, the medial parietal cortex and the TPJ—in subserving the cognitive processes necessary to mentalize about others. Critically, these regions have also been linked to tasks that require projection of oneself away from the here and now and into times or places other than the one currently being inhabited (as during episodic memory, prospection about the future, or spatial navigation). That the same network of brain regions subserves mentalizing suggests that one strategy for understanding other minds is through mental simulation of another person's experience.

However, mentalizing on the basis of self-projection poses a set of difficult cognitive challenges for the human mind, including the need to distinguish individuals for whom one's own mind can reasonably serve as a proxy from those for whom it cannot; the ability to generate rich and accurate representations of one's own hypothetical mental states; and a mechanism by which to suspend one's own experience in order to conjure up the thoughts and feelings of others. Although our understanding of the neural and cognitive bases of these skills has increased considerably in the past decade and a half, fully illuminating the various cognitive processes that endow humans with these abilities continues as a central programme of research in social cognition.

The work described in this paper was supported by NSF BCS 0642448 and NIA R01 AG032780.

REFERENCES

Addis, D. R., Wong, A. T. & Schacter, D. L. 2007 Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. *Neuropsychologica* **45**, 1363–1377. (doi:10.1016/j.neuropsychologia.2006.10.016)

Ames, D. L., Jenkins, A. C., Banaji, M. R. & Mitchell, J. P. 2008 Taking another's perspective increases self-referential neural processing. *Psychol. Sci.* **19**, 642–644. (doi:10.1111/j.1467-9280.2008.02135.x)

APA 1994 *Diagnostic and statistical manual of mental disorders DSM-IV*, 4th edn. Washington, DC: American Psychiatric Association.

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**, 1773–1784. (doi:10.1162/0898929042947928)

Apperly, I. A., Samson, D. & Humphreys, G. W. 2005 Domain-specificity and theory of mind: evaluating neuropsychological evidence. *Trends Cogn. Sci.* **9**, 572–577. (doi:10.1016/j.tics.2005.10.004)

Apperly, I. A., Samson, D., Chiavarino, C., Bickerton, W. L. & Humphreys, G. W. 2006 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**, 300–321. (doi:10.1016/j.cognition.2006.04.012)

Astafiev, S. V., Shulman, G. L., Stanley, C. M., Snyder, A. Z., Van Essen, D. C. & Corbetta, M. 2003 Functional organization of human intraparietal and frontal cortex for attending, looking, and pointing. *J. Neurosci.* **23**, 4689–4699.

Astafiev, S. V., Shulman, G. L. & Corbetta, M. 2006 Visuospatial reorienting signals in the human temporo-parietal junction are independent of response selection. *Eur. J. Neurosci.* **23**, 591–596. (doi:10.1111/j.1460-9568.2005.04573.x)

Baron-Cohen, S. 1995 *Mindblindness: an essay on autism and theory of mind*. Cambridge, MA: MIT Press.

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. (doi:10.1093/brain/awh108)

Bozeat, S., Gregory, C. A., Ralph, M. A. & Hodges, J. R. 2000 Which neuropsychiatric and behavioural features distinguish frontal and temporal variants of frontotemporal dementia from Alzheimer's disease? *J. Neurol. Neurosurg. Psychiatry* **69**, 178–186. (doi:10.1136/jnnp.69.2.178)

Buckner, R. L. & Carroll, D. C. 2007 Self-projection and the brain. *Trends Cogn. Sci.* **11**, 49–57. (doi:10.1016/j.tics.2006.11.004)

Castelli, F., Happé, F., Frith, U. & Frith, C. 2000 Movement and mind: a functional imaging study of perception and interpretation of complex intential movement patterns. *NeuroImage* **12**, 314–325. (doi:10.1006/nimg.2000.0612)

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), 1839–1849. (doi:10.1093/brain/awf189)

Corbetta, M. & Shulman, G. L. 2002 Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **3**, 201–215. (doi:10.1038/nrn755)

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P. & Shulman, G. L. 2000 Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nat. Neurosci.* **3**, 292–297. (doi:10.1038/73009)

Corbetta, M., Tansy, A. P., Stanley, C. M., Astafiev, S. V., Snyder, A. Z. & Shulman, G. L. 2005 A functional MRI study of preparatory signals for spatial location and objects. *Neuropsychologia* **43**, 2041–2056. (doi:10.1016/j.neuropsychologia.2005.03.020)

Davies, M. & Stone, T. (eds) 1995 *Mental simulation: evaluations and applications*, Oxford, UK: Blackwell Publishers.

Fellows, L. K. & Farah, M. J. 2007 The role of ventromedial prefrontal cortex in decision making: judgment under uncertainty or judgment *per se*? *Cereb. Cortex* **17**, 2669–2674. (doi:10.1093/cercor/bhl176)

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. & Frith, C. D. 1995 Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition* **57**, 109–128. (doi:10.1016/0010-0277(95)00692-R)

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U. & Frith, C. D. 2000 Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia* **38**, 11–21. (doi:10.1016/S0028-3932(99)00053-6)

Gallagher, H. L., Jack, A. I., Roepstorff, A. & Frith, C. D. 2002 Imaging the intentional stance in a competitive game. *NeuroImage* **16**(3 Pt 1), 814–821. (doi:10.1006/nimg.2002.1117)

Gilbert, D. T. 2007 *Stumbling on happiness*. New York, NY: Alfred A. Knopf.

Goel, V., Grafman, J., Sadato, N. & Hallett, M. 1995 Modeling other minds. *Neuroreport* **6**, 1741–1746. (doi:10.1097/00001756-199509000-00009)

Goldman, A. I. 1992 Defense of the simulation theory. *Mind Lang.* **7**, 104–119. (doi:10.1111/j.1468-0017.1992.tb00200.x)

Gopnik, A. & Wellman, H. 1992 Why the child's theory of mind is really a theory. *Mind Lang.* **7**, 145–171. (doi:10.1111/j.1468-0017.1992.tb00202.x)

Gopnik, A. & Wellman, H. M. 1994 The theory theory. In *Mapping the mind: domain specificity in cognition and culture* (eds L. A. Hirschfeld & S. A. Gelman), pp. 257–293. New York, NY: Cambridge University Press.

Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. 1998 Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* **74**, 1464–1480. (doi:10.1037/0022-3514.74.6.1464)

Gregory, C., Lough, S., Stone, V., Erzinclioglu, S., Martin, L., Baron-Cohen, S. & Hodges, J. R. 2002 Theory of mind in patients with frontal variant frontotemporal dementia and Alzheimer's disease: theoretical and practical implications. *Brain* **125**(Pt 4), 752–764. (doi:10.1093/brain/awf079)

Grill-Spector, K., Henson, R. & Martin, A. 2006 Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* **10**, 14–23. (doi:10.1016/j.tics.2005.11.006)

Heal, J. 1986 Replication and functionalism. In *Language, mind and logic* (ed. J. Butterfield), pp. 135–150. Cambridge, UK: Cambridge University Press.

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. USA* **105**, 4507–4512. (doi:10.1073/pnas.0708785105)

Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E. & Prigatano, G. P. 2002 Neural correlates of self-reflection. *Brain* **125**(Pt 8), 1808–1814. (doi:10.1093/brain/awf181)

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S. & Heatherton, T. F. 2002 Finding the self? An event-related fMRI study. *J. Cogn. Neurosci.* **14**, 785–794. (doi:10.1162/08989290260138672)

Kennedy, D. P., Redcay, E. & Courchesne, E. 2006 Failing to deactivate: resting functional abnormalities in autism. *Proc. Natl Acad. Sci. USA* **103**, 8275–8280. (doi:10.1073/pnas.0600674103)

Kincade, J. M., Abrams, R. A., Astafiev, S. V., Shulman, G. L. & Corbetta, M. 2005 An event-related functional magnetic resonance imaging study of voluntary and stimulus-driven orienting of attention. *J. Neurosci.* **25**, 4593–4604. (doi:10.1523/JNEUROSCI.0236-05.2005)

Lev-Ran, S., Shamay-Tsoory, S. G., Zangen, A. & Levkovitz, Y. Submitted. Transcranial magnetic stimulation of the ventromedial prefrontal cortex impairs theory of mind learning.

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F. & Kelley, W. M. 2004 Medial prefrontal activity predicts memory for self. *Cereb. Cortex* **14**, 647–654. (doi:10.1093/cercor/bhh025)

Martin, A. & Weisberg, J. 2003 Neural foundations for understanding social and mechanical concepts. *Cogn. Neuropsychol.* **20**, 575–587. (doi:10.1080/02643290342000005)

McCabe, K., Houser, D., Ryan, L., Smith, V. & Trouard, T. 2001 A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl Acad. Sci. USA* **98**, 11 832–11 835. (doi:10.1073/pnas.211415698)

McKhann, G. M., Albert, M. S., Grossman, M., Miller, B., Dickson, D. & Trojanowski, J. Q. 2001 Clinical and pathological diagnosis of frontotemporal dementia: report of the Work Group on Frontotemporal Dementia and Pick's Disease. *Arch. Neurol.* **58**, 1803–1809. (doi:10.1001/archneur.58.11.1803)

Mitchell, J. P. 2005 The false dichotomy between simulation and theory–theory: the argument's error. *Trends Cogn. Sci.* **9**, 363–364. (doi:10.1016/j.tics.2005.06.010)

Mitchell, J. P. 2008 Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cereb. Cortex* **18**, 262–271. (doi:10.1093/cercor/bhm051)

Mitchell, J. P., Macrae, C. N. & Banaji, M. R. 2004 Encoding specific effects of social cognition on the neural correlates of subsequent memory. *J. Neurosci.* **24**, 4912–4917. (doi:10.1523/JNEUROSCI.0481-04.2004)

Mitchell, J. P., Banaji, M. R. & Macrae, C. N. 2005*a* The link between social cognition and self-referential thought in the medial prefrontal cortex. *J. Cogn. Neurosci.* **17**, 1306–1315. (doi:10.1162/0898929055002418)

Mitchell, J. P., Macrae, C. N. & Banaji, M. R. 2005*b* Forming impressions of people versus inanimate objects: social–cognitive processing in the medial prefrontal cortex. *NeuroImage* **26**, 251–257. (doi:10.1016/j.neuroimage.2005.01.031)

Mitchell, J. P., Cloutier, J., Banaji, M. R. & Macrae, C. N. 2006*a* Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Soc. Cogn. Affect. Neurosci.* **1**, 49–55. (doi:10.1093/scan/nsl007)

Mitchell, J. P., Macrae, C. N. & Banaji, M. R. 2006*b* Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron* **50**, 655–663. (doi:10.1016/j.neuron.2006.03.040)

Mitchell, J. P., Ames, D. L. & Gilbert, D. T. Submitted. Neural dissociations supporting intertemporal choice bias.

Moran, J. M., Macrae, C. N., Heatherton, T. F., Wyland, C. L. & Kelley, W. M. 2006 Neuroanatomical evidence for distinct cognitive and affective components of self. *J. Cogn. Neurosci.* **18**, 1586–1594. (doi:10.1162/jocn.2006.18.9.1586)

Salmon, E., Garraux, G., Delbeuck, X., Collette, F., Kalbe, E., Zuendorf, G., Perani, D., Fazio, F. & Herholz, K. 2003 Predominant ventromedial frontopolar metabolic impairment in frontotemporal dementia. *NeuroImage* **20**, 435–440. (doi:10.1016/S1053-8119(03)00346-X)

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**, 499–500. (doi:10.1038/nn1223)

Saxe, R. 2005 Against simulation: the argument from error. *Trends Cogn. Sci.* **9**, 174–179. (doi:10.1016/j.tics.2005.01.012)

Saxe, R. & Kanwisher, N. 2003 People thinking about thinking people: fMRI investigations of theory of mind. *NeuroImage* **19**, 1835–1842. (doi:10.1016/S1053-8119(03)00230-1)

Saxe, R. & Powell, L. J. 2006 It's the thought that counts: specific brain regions for one component of theory of mind. *Psychol. Sci.* **17**, 692–699. (doi:10.1111/j.1467-9280.2006.01768.x)

Saxe, R. & Wexler, A. 2005 Making sense of another mind: the role of the right temporo-parietal junction.

Neuropsychologia **43**, 1391–1399. (doi:10.1016/j.neuropsychologia.2005.02.013)

Schacter, D. L., Addis, D. R. & Buckner, R. L. 2007 Remembering the past to imagine the future: the prospective brain. *Nat. Rev. Neurosci.* **8**, 657–661. (doi:10.1038/nrn2213)

Schmitz, T. W., Kawahara-Baccus, T. N. & Johnson, S. C. 2004 Metacognitive evaluation, self-relevance, and the right prefrontal cortex. *NeuroImage* **22**, 941–947. (doi:10.1016/j.neuroimage.2004.02.018)

Serences, J. T., Shomstein, S., Leber, A. B., Golay, X., Egeth, H. E. & Yantis, S. 2005 Coordination of voluntary and stimulus-driven attentional control in human cortex. *Psychol. Sci.* **16**, 114–122. (doi:10.1111/j.0956-7976.2005.00791.x)

Shamay-Tsoory, S. G., Tomer, R., Goldsher, D., Berger, B. D. & Aharon-Peretz, J. 2004 Impairment in cognitive and affective empathy in patients with brain lesions: anatomical and cognitive correlates. *J. Clin. Exp. Neuropsychol.* **26**, 1113–1127. (doi:10.1080/13803390490515531)

Shamay-Tsoory, S. G., Tibi-Elhanany, Y. & Aharon-Peretz, J. 2006 The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Soc. Neurosci.* **1**, 149–166. (doi:10.1080/17470910600985589)

Shulman, G. L., d'Avossa, G., Tansy, A. P. & Corbetta, M. 2002 Two attentional processes in the parietal lobe. *Cereb. Cortex* **12**, 1124–1131. (doi:10.1093/cercor/12.11.1124)

Spreng, R. N., Mar, R. A. & Kim, A. S. In press. The common neural basis of autobiographical memory, prospection, navigation, theory of mind and the default mode: a quantitative meta-analysis. *J. Cogn. Neurosci.* (doi:10.1162/jocn.2008.21029)

Stone, V. E., Baron-Cohen, S. & Knight, R. T. 1998 Frontal lobe contributions to theory of mind. *J. Cogn. Neurosci.* **10**, 640–656. (doi:10.1162/089892998562942)

Suddendorf, T. & Corballis, M. C. 2007 The evolution of foresight: what is mental time travel, and is it unique to humans? *Behav. Brain Sci.* **30**, 299–313. (doi:10.1017/S0140525X07001975)

Trope, Y. & Liberman, N. 2003 Temporal construal. *Psychol. Rev.* **110**, 401–421. (doi:10.1037/0033-295X.110.3.403)

Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K. & Fink, G. R. 2004 Neural correlates of first-person perspective as one constituent of human self-consciousness. *J. Cogn. Neurosci.* **16**, 817–827. (doi:10.1162/089892904970799)

Wellman, H. M. 1992 *The child's theory of mind*. Cambridge, MA: The MIT Press.

Wheatley, T., Milleville, S. C. & Martin, A. 2007 Understanding animate agents: distinct roles for the social network and mirror system. *Psychol. Sci.* **18**, 469–474. (doi:10.1111/j.1467-9280.2007.01923.x)

Young, L., Camprodon, J. Hauser, M., Pascual-Leone, A. & Saxe, R. Submitted. TMS to the night temporo-parietal junction reduces the role of beliefs in moral judgements.