Neuron 50, 655–663, May 18, 2006 ©2006 Elsevier Inc. DOI 10.1016/j.neuron.2006.03.040

Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others

Jason P. Mitchell,[1,*] C. Neil Macrae,[2] and Mahzarin R. Banaji[1]
[1] Department of Psychology
Harvard University
William James Hall
33 Kirkland Street
Cambridge, Massachusetts 02138
[2] School of Psychology
University of Aberdeen
Aberdeen AB24 2UB
Scotland

Summary

Human social interaction requires the recognition that other people are governed by the same types of mental states—beliefs, desires, intentions—that guide one's own behavior. We used functional neuroimaging to examine how perceivers make mental state inferences when such self-other overlap can be assumed (when the other is similar to oneself) and when it cannot (when the other is dissimilar from oneself). We observed a double dissociation such that mentalizing about a similar other engaged a region of ventral mPFC linked to self-referential thought, whereas mentalizing about a dissimilar other engaged a more dorsal subregion of mPFC. The overlap between judgments of self and similar others suggests the plausibility of "simulation" accounts of social cognition, which posit that perceivers can use knowledge about themselves to infer the mental states of others.

Introduction

Any attempt at understanding the behavior of another person requires a consideration of the rich set of internal mental states that govern what others do and say. Adopting this kind of "intentional stance" requires the recognition that others are mental agents that act primarily on the basis of what they believe, feel, and desire (Dennett, 1987). Moreover, human social cognition includes the added appreciation that others not only experience beliefs, feelings, and desires, but that their mental states are generally comparable to those experienced by oneself, a cognitive ability that appears to distinguish humans from other primates (Tomasello, 1999). This unique awareness that the inner workings of others' minds overlap meaningfully with one's own allows humans to use their own thoughts and feelings as a guide to those of others. Indeed, "simulation" (or "projection") theories of social cognition have posited that perceivers may infer mental states, in part, by assuming that others experience what they themselves would think or feel in a comparable situation (Adolphs, 2002; Davies and Stone, 1995a, 1995b; Gallese and Goldman, 1998; Gordon, 1992; Heal, 1986; Meltzoff and Brooks, 2001; Nickerson, 1999).

Critically, this strategy of basing inferences about others on knowledge about oneself will be useful only to the extent that one can assume that the other is likely to experience the same mental states as oneself. If a target of mentalizing is substantially different from oneself (e.g., somebody from another culture or ethnic group), using self knowledge to inform such inferences may be relatively less useful. As such, simulationist accounts of social cognition suggest that perceivers will mentalize in a different way when the other is perceived to be similar to versus dissimilar from oneself.

Although considerable neuroimaging (Frith and Frith, 1999; Gallagher and Frith, 2003) and neuropsychological (Gregory et al., 2002; Stuss et al., 2001) research has demonstrated that inferences about another's mental states rely on medial aspects of prefrontal cortex (mPFC), few attempts have been made to identify functionally discrete subregions within mPFC that may contribute to differences in the ways mentalizing takes place. Nevertheless, extant data on the neural basis of social cognition provide some preliminary support for the possibility that mPFC contributions to mentalizing may differ according to the perceived overlap between self and other. Although most studies have linked mental state inferences to modulation of dorsal regions of mPFC (Castelli et al., 2000; Fletcher et al., 1995; Gallagher et al., 2000, 2002; Goel et al., 1995; Kumaran and Maguire, 2005; Mitchell et al., 2004, 2005a, 2005b; Saxe and Kanwisher, 2003), some of these studies have also reported the engagement of an additional, ventral mPFC region during social-cognitive tasks (Gallagher et al., 2000, 2002; Kumaran and Maguire, 2005; Mitchell et al., 2005a) that is especially pronounced for one's friends (Kumaran and Maguire, 2005) or similar others (Mitchell et al., 2005a). As yet, little is understood about the specific contributions made by each of these mPFC subregions to social cognition.

However, an important clue regarding the functional distinction between these two regions has recently come from studies of self-referential processing, which have consistently observed selective engagement of ventral mPFC during tasks that require reporting one's own internal states. This ventral mPFC region (typically, within a few millimeters of the axial plane of the genu of the corpus callosum) has been linked to a variety of self-referential tasks (Northoff et al., 2006) such as reporting on one's preferences or personality (Johnson et al., 2002; Kelley et al., 2002; Macrae et al., 2004; Moran et al., 2006; Schmitz et al., 2004; Zysset et al., 2002), reflecting on one's current affective state (Gusnard et al., 2001), or adopting a first-person perspective (Vogeley et al., 2004). Typically, dorsal mPFC activation has not been modulated during self-referencing tasks.

Taken together, the dual observations that ventral mPFC is engaged occasionally during mentalizing and consistently during self-referential processing support the notion that this region may subserve inferences

*Correspondence: mitchell@wjh.harvard.edu

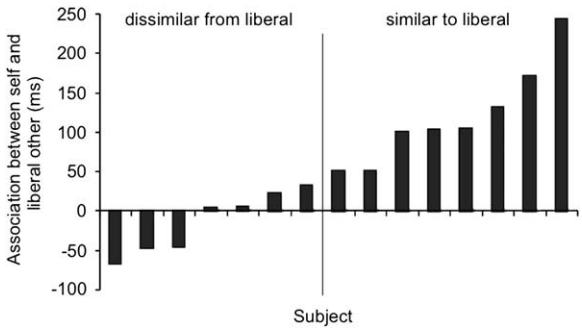

Figure 1. IAT Response Latency Differences for Each Participant

Difference values were obtained by subtracting the mean response latency to trials in the self-with-liberal block from the mean response latency to trials in the self-with-conservative block. Thus, higher values indicate faster responses when first-person pronouns were paired with the liberal target, and thus suggest a stronger association between self and the liberal target. On the basis of these results, participants were retroactively assigned to either the "dissimilar from liberal" or "similar to liberal" participant group.

about others' mental states through simulation, that is, when mental state inferences are informed by perceivers' knowledge about their own feelings and thoughts. If so, simulation theories of social cognition suggest that this region should be specifically engaged for mental state inferences about others perceived to be similar to oneself, since mentalizing on the basis of self knowledge can only take place if another person's internal experience is assumed to be comparable to one's own. As such, this hypothesis suggests an important "division of labor" in the contributions made by different subregions of mPFC to mentalizing. Whereas ventral mPFC may be expected to contribute to mental state inferences about similar others, the dorsal aspects of mPFC—more traditionally associated with mentalizing tasks—should be specifically engaged by mentalizing about dissimilar others, that is, individuals for whom overlap between self and other cannot be assumed.

To test these predictions, we first had participants read descriptions of two unfamiliar individuals who were described as having opposing liberal or conservative sociopolitical views (see Experimental Procedures section). This manipulation was intended to create one target whose social and political views would be similar to those of the participant and another target whose views would be dissimilar. During functional magnetic resonance imaging (fMRI) scanning, participants mentalized about the opinions, likes, and dislikes of each of these two targets and also indicated their own responses on the same set of opinion questions. After scanning, participants completed a version of the Implicit Association Test (IAT) (Greenwald et al., 1998), designed to index how strongly they automatically associated "self" with the liberal target versus the conservative target. Participants also completed a version of the IAT that measured their automatic evaluation of each target (positive versus negative) and, finally, reported their explicit social and political attitudes on a seven-point scale (1 = extremely liberal; 4 = neither liberal nor conservative; 7 = extremely conservative).

Results

Behavioral Data

Participants were significantly faster to make judgments of self (M = 1863 ms) than judgments of others (M = 1927 ms), $t(14) = 2.89$, $p < 0.02$. In addition, participants judged liberal targets more quickly on average than conservative targets (Ms = 1907 and 1947 ms, respectively) $t(14) = 2.21$, $p < 0.05$.

On the two postscanning self-report measures, participants explicitly reported having relatively liberal social (M = 2.93, SD = 1.71) and political (M = 2.80, SD = 1.42) attitudes. As expected, responses on these two explicit questions were highly correlated, $r(14) = 0.93$. Consistent with this liberal outlook, results from the postscanning IAT session indicated that participants were, as a group, significantly faster to categorize trials in self-with-liberal blocks (M = 610 ms) than trials in self-with-conservative blocks (M = 669 ms), $t(14) = 2.67$, $p < 0.02$, demonstrating that, on average, participants more strongly associated self with the liberal than the conservative target. However, sizeable variability across participants was observed in the response latency difference between the two blocks of the IAT, with some participants showing a stronger association between self and the conservative target (SD = 85.9, range = −65.8 to 244.2 ms). We used this variability to segregate participants into two groups (Figure 1). Specifically, based on a median split of their IAT difference score, participants were assigned either to the "similar to liberal" (n = 8, mean IAT difference = 120.2 ms, SD = 63.8) or "dissimilar from liberal" (n = 7, mean IAT difference = −12.8 ms, SD = 38.6) group.

Importantly, this participant group factor did not interact with the reaction time advantage for liberal targets. Participants were equally fast to judge similar targets (i.e., the liberal target for participants in the "similar to liberal" group and the conservative target for participants in the "dissimilar from liberal" group) than to judge dissimilar targets (Ms 1924 and 1929 ms, respectively), $t(14) = 0.49$, $p = 0.63$.

fMRI Data

fMRI data were subjected to a random-effects analysis that included participant group factor (similar to liberal, dissimilar from liberal) in order to identify brain regions that demonstrated a significant interaction of participant group × target (liberal, conservative). This analysis revealed two discrete regions in mPFC. First, a ventral mPFC region (MNI coordinates of peak voxel: 18, 57, 9) was more engaged during judgments of the target with whom participants more strongly associated themselves as measured by the IAT (Figure 2A). That is, participants in the "similar to liberal" group demonstrated greater engagement of ventral mPFC while making mentalizing judgments of the liberal target, whereas participants in the "dissimilar from liberal" group demonstrated greater ventral mPFC engagement while judging the conservative target. Several additional regions—right inferior frontal gyrus, cingulate cortex, and bilateral occipital cortex—were also obtained from this analysis and demonstrated greater activation for similar than dissimilar targets (see Table 1).

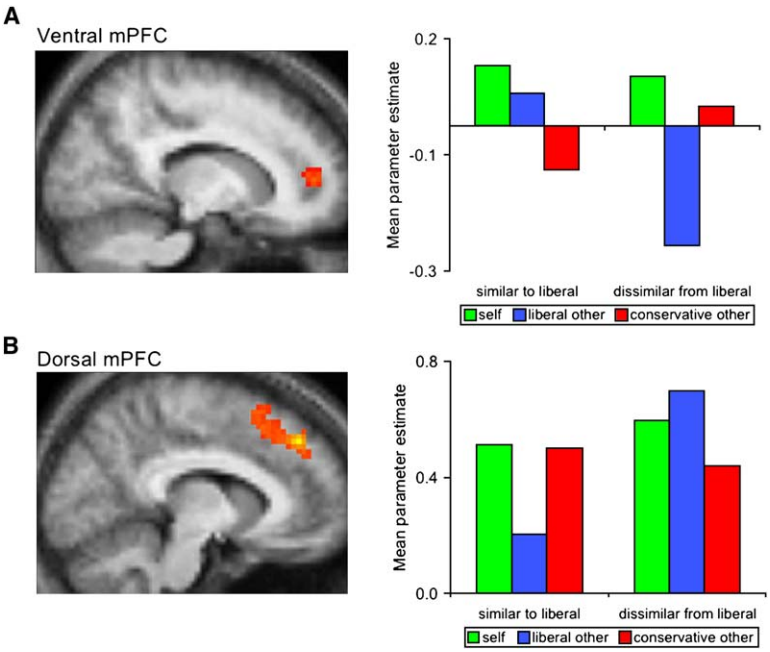

Figure 2. Medial Prefrontal Regions Obtained from Random-Effects Analysis of the Interaction of Participant Group (Similar to Liberal, Dissimilar from Liberal) × Target (Liberal Other, Conservative Other)

(A) A region of ventral mPFC showed greater activation during judgments of the target to whom participants considered themselves to be more similar. For participants who associated self with the liberal target (left set of bars), the response of the ventral mPFC was higher for liberal targets (middle, blue bar) than conservative targets (rightmost, red bar), and no difference was observed for judgments of self (leftmost, green bar) and the liberal target. In contrast, for participants who did not associate with the liberal target (right set of bars), the response of ventral mPFC was higher for conservative than liberal targets, and no difference was observed for judgments of self and the conservative target.

(B) A region of dorsal mPFC showed the opposite pattern of results, that is, greater activation during judgments of the target from whom participants considered themselves to be dissimilar.

In contrast, activation in dorsal mPFC (peak voxel: −9, 45, 42) was greater during judgments of the target with whom participants less strongly associated themselves (Figure 2B). That is, participants in the "similar to liberal" group demonstrated greater engagement of dorsal mPFC while making judgments of the *conservative* target, whereas participants in the "dissimilar from liberal" group demonstrated greater dorsal mPFC engagement while judging the *liberal* target. This region of dorsal mPFC was the only area that showed greater activation for dissimilar than similar targets. Confirming that these two mPFC regions responded differently as a function of target similarity, we observed a highly significant three-way interaction for region (ventral mPFC, dorsal mPFC) × target (liberal, conservative) × participant group (similar to liberal, dissimilar from liberal), $F(1,13) = 26.06$, $p < 0.0002$.

Table 1. Coordinates of Peak Activations and Percent Signal Change for Regions Demonstrating a Significantly Different BOLD Response for Similar and Dissimilar Targets

Anatomical Label	x	y	z	Similar	Dissimilar
Similar > dissimilar					
Ventral mPFC	18	57	9	0.18	−0.11
R inferior frontal gyrus	51	3	24	0.85	0.39
Cingulate cortex	−3	3	36	0.95	0.45
R occipital cortex	12	−66	−6	0.89	0.43
	9	−90	24	0.77	0.24
L occipital cortex	−24	−66	24	0.18	−0.08
Dissimilar > similar					
Dorsal mPFC	−9	45	42	0.57	0.82

Peak activations are reported for each region in the Montreal Neurological Institute stereotaxic space. The two rightmost columns present percent signal change as a function of the perceived similarity of targets. For the purposes of reporting percent signal change in each of these regions, the liberal target was considered "similar" and the conservative target was considered "dissimilar" for participants in the "similar-to-liberal" group (and vice versa for participants in the "dissimilar-from-liberal" group). mPFC = medial prefrontal cortex; R = right; L = left.

These findings were further supported by correlational analyses that capitalized on the full range of variability in participants' IAT results. As displayed in Figure 3A, the activation in ventral mPFC during judgments of the liberal target (relative to the conservative target) was significantly correlated with the extent to which a participant associated self with the liberal target on the postscanning IAT, $r(14) = 0.54$, $p < 0.04$. That is, the more a participant associated self with the liberal target (as measured by the IAT), the greater the difference in ventral mPFC activation during judgments of the liberal target relative to judgments of the conservative target. The inverse pattern was observed in dorsal mPFC (Figure 3B), such that the relative activity during judgments of the liberal target was *negatively* correlated with the extent to which a participant associated self with the liberal other on the IAT, $r(14) = -0.72$, $p < 0.003$. That is, the *less* a participant associated self with the liberal target, the greater the difference in dorsal mPFC activation during judgments of the liberal target (relative to judgments of the conservative target). In other words, whereas the response in ventral mPFC tracked how similar the participants considered themselves to a target, the response in dorsal mPFC tracked with how *dissimilar* participants considered themselves from a target.

Finally, we examined the overlap between self and other through a separate analysis in which judgments of others were reconditionalized on the basis of the distance between response to other and response for self. We reasoned that participants were relatively likely to have used knowledge about their own opinions and predilections when making the same response for a target as for themselves (e.g., judging that a target looked forward to returning home for Thanksgiving as much as they themselves did). Accordingly, judgments of the two targets were segregated into (1) those for which

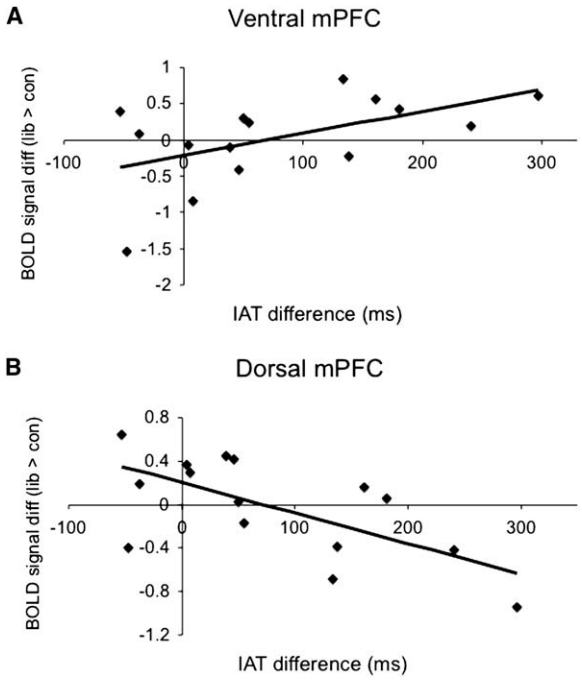

Figure 3. Scatter Plots Displaying the Relation between BOLD Response and IAT Difference Score in mPFC for Each Participant

The x axes display the difference of the parameter estimate associated with liberal trials minus the parameter estimate associated with conservative trials (thus, greater values indicate greater activation during judgments of the liberal target). The y axes display the IAT difference score calculated the same way as described in Figure 1 (higher values indicate stronger association with the liberal target). Each point represents one participant. (A) In ventral mPFC, the difference in BOLD response between liberal and conservative targets was positively correlated with the strength of association between self and the liberal target. That is, participants who associated themselves most with the liberal target also showed the largest relative BOLD response in this ventral region during judgments of the liberal target. (B) In contrast, in dorsal mPFC the difference in BOLD response between liberal and conservative targets was negatively correlated with the strength of association between self and the liberal target. That is, participants who associated themselves most with the liberal target also showed the smallest relative BOLD response in this dorsal region during judgments of the liberal target.

the participant made the same judgment of the other as for self (other-same trials) and (2) those for which the participant made a different judgment for the other than for self (other-different trials). As displayed in Figure 4, activation of ventral mPFC was nearly identical during judgments of self and other-same judgments, [$t(14) = 0.29$, *ns*], but both of these trial types engaged ventral mPFC significantly more than other-different judgments (both t values > 3.12, both p values < 0.01). This additional analysis provides further support for the notion that the ventral mPFC may be engaged when using knowledge about oneself to mentalize about others. Although a numerical trend in the opposite direction was observed in dorsal mPFC, the comparable analysis in this region did not yield significant results.

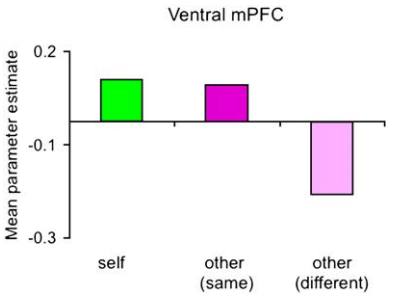

Figure 4. Responses in Ventral mPFC as a Function of the Overlap between Behavioral Judgments of Self and Other

Response in ventral mPFC during self trials, "other-same" targets that were given the same response as for oneself, and "other-different" targets that were given a different response from oneself. Whereas the response in this region was similar for self and other-same trials, the response to both of these trial types was significantly greater than for other-different trials.

Discussion

These results begin the process of identifying functionally discrete subregions of mPFC, suggesting that distinct dorsal and ventral sections of the medial wall of prefrontal cortex subserve social-cognitive processing as a function of how similar another person is perceived to be to oneself. A ventral region of mPFC—overlapping with earlier studies that have implicated this region in self-referential processing (Gusnard et al., 2001; Johnson et al., 2002; Kelley et al., 2002; Macrae et al., 2004; Schmitz et al., 2004; Vogeley et al., 2004; Zysset et al., 2002)—was more strongly engaged during judgments about the potential mental states of others perceived to be similar to versus dissimilar from oneself. In contrast, a more dorsal region of mPFC—more consistent with the bulk of studies that have identified regions activated by mentalizing tasks—demonstrated the opposite effect of being more strongly engaged during judgments about the potential mental states of others perceived to be dissimilar from oneself. This double dissociation was obtained despite measuring target similarity through an unrelated behavioral measure that was administered postscanning and which simply assessed the relative speed with which participants associated first-person pronouns with photographs of the two target individuals. Moreover, secondary analyses demonstrated that ventral mPFC was similarly engaged for judgments about oneself and judgments of others who were thought to share the same opinion as the participant on a particular question.

This functional division has not been evident in the sizeable number of studies that have linked mPFC activity to social cognition, which have generally linked mentalizing to relatively dorsal aspects of mPFC. By and large, however, these studies have asked participants to mentalize about targets that they were unlikely to perceive as similar to self, such as historical figures (Goel et al., 1995), cartoon and story characters (Fletcher et al., 1995; Gallagher et al., 2000; Saxe and Kanwisher, 2003), and abstract shapes (Castelli et al., 2000). As such, these earlier studies may have inadvertently isolated a subregion of mPFC that contributes specifically to mentalizing about relatively dissimilar others. Interestingly, the small number of studies that have reported ventral mPFC engagement during mentalizing tasks have typically included targets that were known to be similar to participants (Kumaran and Maguire, 2005;

Mitchell et al., 2005a) or else prompted mentalizing in the context of computer-based games in which players may adopt the first-person perspective of their opponents (Gallagher et al., 2002; McCabe et al., 2001). Manipulating the similarity of targets within a single experiment provided the opportunity to observe this dissociation between ventral and dorsal subregions of mPFC.

A recent study by Saxe and Wexler (2005) also included a manipulation that would have likely altered the perceived similarity of targets. In this study, participants made mentalizing judgments about targets that were described as having social and cultural backgrounds that were either "familiar" or "foreign." Intriguingly, these researchers report that activation in a region of mPFC was marginally greater ($p < 0.10$) for targets with foreign than familiar backgrounds, suggesting that this region responds more to dissimilar than similar others. However, although the "representative" region displayed by the authors from a single participant suggests that this effect was observed in dorsal aspects of mPFC (consistent with the current results), these authors do not report analyses that would permit a more detailed comparison between their data and the results of the current study.

In an earlier paper (Mitchell et al., 2005a), we demonstrated a similar modulation of ventral mPFC activity as a function of perceived similarity between self and other. In that paper, participants either made mentalizing ("how pleased was this person to have her [or his] photograph taken?") or nonmentalizing ("how symmetrical is this face?") judgments about a large number of targets. After scanning, participants saw each of the faces a second time and were asked to indicate how similar or dissimilar the target was to themselves. These postscanning ratings were used to identify a region of the ventral mPFC in which activity correlated with these subsequent similarity ratings (higher activity for more similar targets), but only for targets initially encountered as part of the mentalizing task. Although this earlier study is important for demonstrating that perceived self-other similarity is only relevant when making social-cognitive judgments about another's mental states, the current study extends this earlier work in several critical ways. First, the manipulation of the targets' political views in the current study better specified the way in which perceivers saw themselves as similar or dissimilar from each target (in the previous study, participants were free to use whatever dimension of similarity they saw fit, whereas here we fixed sociopolitical attitudes as the relevant dimension of similarity). More importantly, the current results reveal a full double dissociation between ventral and dorsal mPFC, thereby strengthening the empirical case for a division of labor in mPFC contributions to mentalizing. This double dissociation also rules out the possibility of an item effect, whereby some targets were simply more likely to be seen as similar by all participants, since the same targets engaged different subregions of mPFC as a function of the between-subject group to which a participant was later assigned.

These data suggest the plausibility of simulation accounts of social cognition, which posit that an understanding of another's mind can be informed through first-person experience of one's own. Several variants of simulation have been discussed in the context of social cognition. The term was originally introduced by theorists to refer to circumstances in which one infers another's thoughts or feelings by mentally imagining oneself in the same situation as a target (Adolphs, 2002; Davies and Stone, 1995a, 1995b; Gallese and Goldman, 1998; Gordon, 1992; Heal, 1986; Meltzoff and Brooks, 2001; Nickerson, 1999). Recently, the idea of simulation has been expanded to include the related, but distinct, phenomenon of mentalizing by placing oneself in the same bodily state as a target, for example, by displaying the same facial expression, potentiating the same motor response, or activating the same brain regions as another person (Gallese and Goldman, 1998). Although these recently introduced "resonance" (or "mirror") versions of simulation provide a useful account for how perceivers might infer the emotional experience or predict the actions of others with whom they are interacting, they are necessarily limited to those situations in which visual or auditory cues about the target's mental states are available (i.e., an individual can directly see or hear the target of mentalizing). Moreover, such resonance-based simulation is useful only when perceptual cues successfully communicate information about the relevant internal states of the targets; certain mental states—such as the kinds of attitudes and stable predilections that our participants were asked to judge from prelearned knowledge about a target—cannot be inferred on the basis of mirroring another's current physical states. Instead, we suggest that the results of the current study are consistent with the possibility that perceivers make selective use of simulation in the original sense, plumbing their own possible—but not necessarily concurrently experienced—thoughts and feelings for clues to those of others.

Of course, targets perceived to be highly similar to oneself are also likely to differ from dissimilar others in a number of other important ways. We have long known that familiar objects are evaluated more positively than novel ones (Zajonc, 1968), an effect that extends to other people (Monin, 2003). Consistent with this view, on a second postscanning IAT designed to measure associations between each target and affectively laden words (paradise, cockroach), participants demonstrated greater positivity toward the similar relative to the dissimilar target, $F(1,13) = 4.91$, $p < 0.05$; for example, participants in the "similar to liberal" group responded more quickly to trials within blocks in which the liberal target was paired with positive words than negative words. Unsurprisingly, correlational analyses revealed that the two IAT measures were significantly related, $r(14) = 0.56$, $p < 0.05$. Moreover, activity in ventral mPFC was correlated with the evaluation IAT in the same way that it was correlated with the identity IAT; that is, the strength of activation to the liberal (relative to the conservative) target was significantly related to the strength of positivity to the liberal target, $r(14) = 0.58$, $p < 0.05$. The comparability of the relation between ventral mPFC activation and both IAT measures suggests that it could be positivity—rather than similarity—that differentiates ventral from dorsal mPFC contributions to mentalizing. However, this possibility is challenged by a failure to obtain any significant relation

between evaluation and activation in dorsal mPFC. That is, unlike the strong inverse correlation obtained between dorsal mPFC activity and the identity IAT, no significant relation with the evaluation IAT was observed in dorsal mPFC, $r(14) = -0.28$, *ns*. Moreover, the alternative interpretation for the observed mPFC dissociation is undermined by recent demonstrations (Moran et al., 2006) that, while ventral mPFC is differentially engaged by the degree of self-relevant processing, its activity does not correlate with affective/emotion-based processing per se (which seems to engage a more posterior and ventral region of anterior cingulate cortex). Finally, we note that the current results are inconsistent with the view that ventral mPFC activity tracks with affective processing, since the correlation we observed was specific for the relative degree of *positivity* that was demonstrated toward a target individual, not the overall amount of affect associated with the person. As such, the overall pattern of results is more consistent with the view that ventral and dorsal mPFC are differentially sensitive to the perceived similarity between oneself and the target of mentalizing than differences in positive and negative evaluation of targets. However, cleaving the tight relation between the similarity of a target and the positivity felt toward that person remains a significant challenge in research of this kind.

Recent work by Völlm et al. (2006) has suggested that ventral mPFC may be particularly engaged when perceivers make inferences about the affective aspects of another person's mental states (e.g., feelings, desires, and motivations), whereas dorsal mPFC subserves inferences about both affective as well as "colder," more cognitive mental states, such as beliefs and knowledge. Interestingly, the extant behavioral studies that most strongly support simulationist accounts of mentalizing have almost uniformly examined affective mentalizing (Mitchell, 2005), suggesting a link between these authors' proposal that ventral mPFC may contribute specifically to mentalizing about affective states and the current suggestion that this region contributes to mentalizing via self reference. For example, Niedenthal and colleagues (Niedenthal et al., 2001; 2000) have demonstrated that observers' own emotional state strongly colors their judgments of others' emotions (e.g., sad observers more readily perceive sadness in ambiguous facial displays than happy observers), except when prevented from spontaneously mimicking the facial expression of the person being judged. Likewise, support for the hypothesis that perceivers should reserve self-referential processes for mentalizing about similar others—and that this effect might be especially pronounced for affective mentalizing—has come from demonstrations that observers assume that outgroup members do not experience the same depth of emotional experience that they do themselves (Demoulin et al., 2004; Vaes et al., 2003) and more readily project their own goals and predilections onto similar targets than dissimilar ones (Ames, 2004). As in this earlier research, participants in the current study were specifically asked about the likes, dislikes, and attitudinal opinions of targets.

In contrast, much of the data marshalled against the possibility that humans infer each others' mental experience through the exclusive use of simulation has investigated mentalizing about more cognitive mental states, such as beliefs and knowledge. In her recent review of behavioral evidence that weighs against simulation accounts, Saxe (2005) points out a number of situations in which mentalizing performance is unlikely to be self-referential. Each of these is limited to mentalizing about beliefs and knowledge. For example, Ruffman (1996) has demonstrated that young children do not appear to engage in simulation when inferring another person's knowledge; instead, children appear to assume that a respondent who is ignorant of the color of a randomly chosen bead will systematically answer incorrectly, which is inconsistent with the simulationist prediction that perceivers should know that respondents would actually be agnostic about the actual state of affairs. This formulation suggests that the use of self-reference to understand others' mental states may depend, in part, on the particular mental states that need to be inferred (Mitchell, 2005); whereas affective mentalizing may draw heavily on self-reference, mentalizing about beliefs and knowledge may rely on a different set of cognitive processes (Saxe and Kanwisher, 2003). That such a dichotomy between affective and cognitive mentalizing may exist is a contribution to our understanding of social cognition made uniquely by recent neuroimaging research, and suggests an important avenue for future investigations into the apparent link between self-reference, affective mentalizing, and the ventral mPFC.

Lastly, we point out the potential relevance of this work for understanding the nature of outgroup stereotyping and prejudice. Earlier social psychological research has suggested that perceivers tend to "infrahumanize" members of other groups by proving unwilling to acknowledge that outgroup members can experience certain higher-order mental states, such as the second-order emotions of love and guilt (Demoulin et al., 2004; Vaes et al., 2003). To the extent that members of a social group other than one's own are viewed as dissimilar from oneself, the current results suggest that perceivers may actively deploy a different set of social-cognitive processes when considering the mental states of someone of a different race or ethnicity than a member of one's own ingroup. As such, prejudice may arise in part because perceivers assume that outgroup members' mental states do not correspond to their own and, accordingly, mentalize in a non-self-referential way about the minds of people from different groups. Without a self-referential basis for mentalizing about outgroup members, perceivers may rely heavily on precomputed judgments—such as stereotypes—to make mental state inferences about very dissimilar others. This view suggests that a critical strategy for reducing prejudice may be to breach arbitrary boundaries based on social group membership by focusing instead on the shared similarity between oneself and outgroup members.

Experimental Procedures

Participants

Participants were 15 (nine male) right-handed, native English speakers with no history of neurological problems (mean age, 24.4 years; range, 21–29). All participants were undergraduate or graduate students at universities in the Boston area. Informed consent

was obtained in a manner approved by the Human Studies Committee of the Massachusetts General Hospital.

Stimuli and Behavioral Procedure

Participants were told that we were investigating their ability to extrapolate another person's opinions, likes, and dislikes from a small amount of information about that person. Prior to scanning, participants were introduced to two target individuals, represented by face photographs downloaded from an Internet dating site. Each target face was accompanied by a short descriptive paragraph intended to create a sense of similarity between the participant and one target and dissimilarity between the participant and the other target. For one target (randomly determined for each participant), this paragraph described the person as having liberal sociopolitical views and participating in activities typical of many students at Northeast liberal arts colleges. For the other target, the paragraph described the person as a fundamentalist Christian with conservative political and social views who participated avidly in a variety of events sponsored by religious and Republican organizations at a Midwest university. Participants were told that we had generated these descriptions from information provided by the target individuals on an Internet dating site, and that we would subsequently ask them to use what they learned about each target to judge that person's opinions, likes, and dislikes. Order of presentation (liberal target first; conservative target first) was randomized across participants, and the two targets were matched to the sex of the participant (i.e., male participants read about two men; female participants read about two women). Participants were given as much time as needed to read about each of the two targets.

During the subsequent scanning phase, participants judged how likely targets were to agree with each of 66 opinion questions. Each trial began with the appearance of a four-point response scale (anchored by "1 = definitely not" and "4 = definitely") below one of three targets: (1) the photograph of the liberal target used during learning, (2) the photograph of the conservative target used during learning, or (3) a chalk outline of a person's head with the word "me" written inside (used to indicate the subject himself or herself). After 1 s, an opinion question appeared between the face and the response scale, and participants were asked to use the scale either to judge how likely the target was to agree with the question or, on trials when the target was oneself, to indicate how much they personally agreed with the question. Questions referred to a wide range of personal and societal issues (e.g., "to look forward to going home for Thanksgiving?"; "to enjoy having a roommate from a different country?"; "to drive a small car entirely for environmental reasons?"; "to think that European films are generally better than the ones made in Hollywood?"; "to believe that cultural diversity should be an important national issue?"; etc.). The face, question, and scale remained onscreen together for an additional 3 s, during which time participants were obliged to make their response. To increase their engagement with the task, participants were told that we knew about the targets' actual opinions and predilections from questionnaires that they had filled out at the dating web site, and that we were interested in how accurately participants could infer those characteristics in each target; in actual fact, all the information presented was generated specifically for this experiment. Target photographs were resized to a width between 5.54 and 6.35 cm and a height between 6.77 and 8.12 cm. To optimize estimation of the event-related fMRI response, trials were intermixed in a pseudorandom order and separated by a variable interstimulus interval (500—7500 ms) (Dale, 1999), during which participants passively viewed a fixation crosshair.

Approximately 10 min after completing the last functional run, participants completed a version of the Implicit Association Test (Greenwald et al., 1998) that was designed to measure how similar they perceived themselves to be to each of the two targets. The IAT assesses the conceptual association between two classes of stimuli by measuring differences in the speed with which participants can make the same behavioral response to exemplars from two categories (e.g., pressing the same button for pictures of snakes and positively-valenced words compared to pressing the same button for snakes and negatively-valenced words). On each trial of the identity IAT used in the current study, participants categorized an exemplar from one of four categories of stimuli: a photo of the liberal target, a photo of the conservative target, a first-person pronoun (me, mine, or my), or a third-person pronoun (they, theirs, or they). In one block of trials, words related to self (i.e., first-person pronouns) required the same behavioral response as photos of the liberal target (self-with-liberal block); specifically, participants responded with the left key ("d") for first-person pronouns or the photo of the liberal target and responded with the right key ("k") for third-person pronouns or the photo of the conservative target. In another block of trials, words related to self required the same response as photos of the conservative target (self-with-conservative block); in this block, participants responded with the left key for first-person pronouns or the photo of the conservative target and the right key for third-person pronouns or the photo of the liberal target. IAT data were coded in the direction of association between self and the liberal target, that is, as the difference in mean response latency to trials in the self-with-conservative block minus trials in the self-with-liberal block. As such, higher IAT difference scores indicate greater association between self and the liberal target. Each block consisted of 60 trials (15 each of the four trial types), and block order (self-with-liberal first; self-with-conservative first) was randomized across participants.

After this task, participants completed a second IAT that was designed to assess the strength of their evaluations (positivity versus negativity) toward each target. The procedure for this evaluation IAT was identical to that of the identity IAT above, except that first- and third-person pronouns were replaced with positive and negative words. The strength of positivity toward the liberal target was calculated as the difference in average response latency to trials within the negative-with-liberal block minus positive-with-liberal block.

Following conventional treatment of reaction time data (Fazio, 1990; Ratcliff, 1993), IAT data were log-transformed prior to statistical analysis. A trial was excluded from analysis if either (1) the participant incorrectly categorized the stimulus or (2) its response latency was more than three standard deviations from the participant's own mean. These criteria resulted in the exclusion of 6.0% of IAT trials. Reanalysis of these data using the IAT scoring procedures recommended by Greenwald et al. (2003) did not qualify any of the reported results.

Finally, participants were asked two questions regarding their explicit sociopolitical attitudes: (1) "On a scale from 1–7, how socially liberal or conservative would you say you are?" and (2) "On a scale from 1–7, how politically liberal or conservative would you say you are?" The two questions were presented in random order, and participants responded to each using a seven-point Likert-type scale, anchored by 1 = extremely liberal; 4 = neither liberal nor conservative; 7 = extremely conservative. These self-report measures were included only to assess the explicit attitudes of our participant population and were not used to qualify fMRI results. Almost all participants expressed liberal attitudes; indeed, only three participants used a point above the midline to describe their sociopolitical attitudes. This lack of variability precluded the use of this self-report measure as a covariate for fMRI data analysis.

Imaging Procedure

Imaging was conducted using a 1.5 Tesla Siemens Sonata scanner. We first collected a high-resolution T1-weighted structural scan (MP-RAGE) followed by three functional runs of 172 volume acquisitions (26 axial slices; 5 mm thick; 1 mm skip). Functional scanning used a gradient-echo echo-planar pulse sequence (TR = 2 s; TE = 35 ms; 3.75 × 3.75 in-plane resolution). Using PsyScope software (Cohen et al., 1993) for Macintosh OS X, stimuli were projected onto a screen at the end of the magnet bore that participants viewed by way of a mirror mounted on the head coil. A pillow and foam cushions were placed inside the head coil to minimize head movements.

fMRI data were preprocessed and analyzed using SPM99 (Wellcome Department of Cognitive Neurology, London, UK). First, functional data were time corrected for differences in acquisition time between slices for each whole-brain volume and realigned to correct for head movement. Functional data were then transformed into a standard anatomical space (3 mm isotropic voxels) based on the ICBM 152 brain template (Montreal Neurological Institute). Normalized data were then spatially smoothed (8 mm full-width-at-half-maximum [FWHM]) using a Gaussian kernel. Trials were conditionalized as a function of which target was being judged (liberal,

conservative). Statistical analyses were performed using the general linear model in which the event-related design was modeled using a canonical hemodynamic response function, its temporal derivative, and additional covariates of no interest (a session mean and a linear trend). This analysis was performed individually for each participant, and contrast images for each participant were subsequently entered in a second-level analysis which treated participants as a random effect and used participant group (similar to liberal, dissimilar from liberal) as a between-subject factor. As such, comparisons of interest identified brain regions demonstrating a two-way interaction of target (liberal, conservative) × participant group (similar to liberal, dissimilar from liberal). Peak coordinates were identified using a statistical criterion of 130 or more contiguous voxels at a voxel-wise threshold of $p < 0.01$. This cluster size was selected on the basis of a Monte Carlo simulation (S. Slotnick, Boston College) of our brain volume that found that this cluster extent cutoff provided an experiment-wise threshold of $p < 0.05$, corrected for multiple comparisons.

In a secondary analysis, trials were reconditionalized as (1) self, i.e., judgments about oneself; (2) other-same, i.e., either a liberal or conservative target on which, for a particular opinion question, the participant made the same response to self and other; and (3) other-different, i.e., either a liberal or conservative target on which the participant made a different response for self and other. Region-of-interest analyses of the difference between these trial types were conducted on the mPFC regions observed from the primary analysis using analysis-of-variance procedures on the parameter estimates associated with each trial type.

Although analysis of behavioral data suggested that reaction time was highly unlikely to account for any of the neuroimaging results we report, we also performed a supplementary analysis in which participants' trial-by-trial reaction time during the mentalizing task was used as a covariate in analysis of fMRI data. Specifically, we fit a model in which the reaction time associated with each trial was included as a linear parametric modulator. A random-effects analysis was then used to reveal brain regions in which BOLD response was correlated with reaction time: this analysis revealed a number of areas in visual cortex and language-related regions (which frequently covary with reading time), but did not demonstrate any regions in mPFC. More critically, we also directly examined how well reaction time predicted the BOLD response in our particular regions of interest. This analysis was conducted on the parameter estimate associated with the linear relation between reaction time and BOLD response in each of the regions observed from the primary random-effects analyses. Consistent with the above analysis, reaction time was significantly related to the BOLD response only in occipital cortex, but was not related to the response in either ventral mPFC ($p = 0.33$) or dorsal mPFC ($p = 0.21$).

Acknowledgments

We thank E. Aminoff, D. Carney, B. Hughes, A. Jenkins, H. Kober, E. Mela, K. Olson, and S. Gershman for advice and assistance. fMRI data were collected at the Athinoula A. Martinos Center for Biomedical Imaging, which is supported by grant P41RR14075 from the National Center for Research Resources and by a grant from the Mental Illness and Neuroscience Discovery (MIND) Institute. J.P.M. was supported by a postdoctoral National Research Service Award (NRSA). C.N.M. was supported by a Royal Society-Wolfson Fellowship.

Received: January 12, 2006
Revised: March 6, 2006
Accepted: March 30, 2006
Published: May 17, 2006

References

Adolphs, R. (2002). Neural systems for recognizing emotion. Curr. Opin. Neurobiol. *12*, 169–177.

Ames, D.R. (2004). Inside the mind reader's tool kit: projection and stereotyping in mental state inference. J. Pers. Soc. Psychol. *87*, 340–353.

Castelli, F., Happé, F., Frith, U., and Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intential movement patterns. Neuroimage *12*, 314–325.

Cohen, J.D., MacWhinney, B., Flatt, M., and Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. Behav. Res. Methods Instrum. Comput. *25*, 257–271.

Dale, A.M. (1999). Optimal experimental design for event-related fMRI. Hum. Brain Mapp. *8*, 109–114.

Davies, M., and Stone, T. (1995a). Folk Psychology: The Theory of Mind Debate (Oxford, UK: Blackwell Publishers).

Davies, M., and Stone, T. (1995b). Mental Simulation: Evaluations and Applications (Oxford, UK: Blackwell Publishers).

Demoulin, S., Torres, R.R., Perez, A.R., Vaes, J., Paladino, M.P., Gaunt, R., Pozo, B.C., and Leyens, J.-P. (2004). Emotional prejudice can lead to infra-humanisation. In European Review of Social Psychology, W. Stroebe and M. Hewstone, eds. (Hove, England: Psychology Press/Taylor & Francis), pp. 259–296.

Dennett, D.C. (1987). The Intentional Stance (Cambridge, MA: MIT Press).

Fazio, R.H. (1990). A practical guide to the use of response latency in social psychological research. In Research Methods in Personality and Social Psychology, C. Hendrick and M.S. Clark, eds. (Thousand Oaks, CA: Sage Publications), pp. 74–97.

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

Frith, C.D., and Frith, U. (1999). Interacting minds–A biological basis. Science *286*, 1692–1695.

Gallagher, H.L., and Frith, C.D. (2003). Functional imaging of 'theory of mind'. Trends Cogn. Sci. *7*, 77–83.

Gallagher, H.L., Happé, F., Brunswick, N., Fletcher, P.C., Frith, U., and 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.

Gallagher, H.L., Jack, A.I., Roepstorff, A., and Frith, C.D. (2002). Imaging the intentional stance in a competitive game. Neuroimage *16*, 814–821.

Gallese, V., and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. Trends Cogn. Sci. *2*, 493–501.

Goel, V., Grafman, J., Sadato, N., and Hallett, M. (1995). Modeling other minds. Neuroreport *6*, 1741–1746.

Gordon, R.M. (1992). Folk psychology as simulation. Mind and Language *1*, 158–171.

Greenwald, A.G., McGhee, D.E., and Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. J. Pers. Soc. Psychol. *74*, 1464–1480.

Greenwald, A.G., Nosek, B.A., and Banaji, M.R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. J. Pers. Soc. Psychol. *85*, 197–216.

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

Gusnard, D.A., Akbudak, E., Shulman, G.L., and Raichle, M.E. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. Proc. Natl. Acad. Sci. USA *98*, 4259–4264.

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

Johnson, S.C., Baxter, L.C., Wilder, L.S., Pipe, J.G., Heiserman, J.E., and Prigatano, G.P. (2002). Neural correlates of self-reflection. Brain *125*, 1808–1814.

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

Kumaran, D., and Maguire, E.A. (2005). The human hippocampus: cognitive maps or relational memory? J. Neurosci. *25*, 7254–7259.

Macrae, C.N., Moran, J.M., Heatherton, T.F., Banfield, J.F., and Kelley, W.M. (2004). Medial prefrontal activity predicts memory for self. Cereb. Cortex *14*, 647–654.

McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. Proc. Natl. Acad. Sci. USA *98*, 11832–11835.

Meltzoff, A.N., and Brooks, R. (2001). "Like me" as a building block for understanding other minds: Bodily acts, attention, and intention. In Intentions and Intentionality: Foundations of Social Cognition, B.F. Malle, L.J. Moses, and D.A. Baldwin, eds. (Cambridge, MA: MIT Press), pp. 171–191.

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

Mitchell, J.P., Macrae, C.N., and Banaji, M.R. (2004). Encoding specific effects of social cognition on the neural correlates of subsequent memory. J. Neurosci. *24*, 4912–4917.

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

Mitchell, J.P., Macrae, C.N., and Banaji, M.R. (2005b). Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. Neuroimage *26*, 251–257.

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. J. Pers. Soc. Psychol. *85*, 1035–1048.

Moran, J.M., Macrae, C.N., Heatherton, T.F., Wyland, C.L., and Kelley, W.M. (2006). Neuroanatomical evidence for distinct cognitive and affective components of self. J. Cogn. Neurosci., in press.

Nickerson, R. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. Psychol. Bull. *125*, 737–759.

Niedenthal, P.M., Halberstadt, J.B., Margolin, J., and Innes-Ker, Ö.H. (2000). Emotional state and the detection of change in facial expression of emotion. Eur. J. Soc. Psychol. *30*, 211–222.

Niedenthal, P.M., Brauer, M., Halberstadt, J.B., and Innes-Ker, Ö.H. (2001). When did her smile drop: Facial mimicry and the influences of emotional state on the detection of change in emotional expression. Cogn. Emotion *15*, 853–864.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., and Panksepp, J. (2006). Self-referential processing in our brain - A meta-analysis of imaging studies of the self. Neuroimage., in press. Published online February 5, 2006. 10.1016/j.neuroimage.2005.12.002.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. Psychol. Bull. *114*, 510–532.

Ruffman, T. (1996). Do children understand the mind by means of simulation or a theory? Evidence from their understanding of inference. Mind and Language *11*, 387–414.

Saxe, R. (2005). Against simulation: The argument from error. Trends Cogn. Sci. *9*, 174–179.

Saxe, R., and Kanwisher, N. (2003). People thinking about thinking people: fMRI investigations of theory of mind. Neuroimage *19*, 1835–1842.

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

Schmitz, T.W., Kawahara-Baccus, T.N., and Johnson, S.C. (2004). Metacognitive evaluation, self-relevance, and the right prefrontal cortex. Neuroimage *22*, 941–947.

Stuss, D.T., Gallup, G.G., Jr., and Alexander, M.P. (2001). The frontal lobes are necessary for 'theory of mind'. Brain *124*, 279–286.

Tomasello, M. (1999). The Cultural Origins of Human Cognition (Cambridge, MA: Harvard University Press).

Vaes, J., Paladino, M.P., Castelli, L., Leyens, J.-P., and Giovanazzi, A. (2003). On the behavioral consequences of infrahumanization: The implicit role of uniquely human emotions in intergroup relations. J. Pers. Soc. Psychol. *85*, 1016–1034.

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

Völlm, B.A., Taylor, A.N., Richardson, P., Corcoran, R., Stirling, J., McKie, S., Deakin, J.F., and Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. Neuroimage *29*, 90–98.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. J. Pers. Soc. Psychol. *9*, 1–27.

Zysset, S., Huber, O., Ferstl, E., and von Cramon, D.Y. (2002). The anterior frontomedian cortex and evaluative judgment: an fMRI study. Neuroimage *15*, 983–991.