Repetition suppression of ventromedial prefrontal activity during judgments of self and others

Adrianna C. Jenkins*†, C. Neil Macrae‡, and Jason P. Mitchell*

*Department of Psychology, Harvard University, Cambridge, MA 02138; and ‡School of Psychology, University of Aberdeen, Aberdeen AB24 2UB, United Kingdom

Edited by Edward E. Smith, Columbia University, New York, NY, and approved January 31, 2008 (received for review September 15, 2007)

One useful strategy for inferring others' mental states (i.e., mentalizing) may be to use one's own thoughts, feelings, and desires as a proxy for those of other people. Such self-referential accounts of social cognition are supported by recent neuroimaging observations that a single brain region, ventromedial prefrontal cortex (vMPFC), is engaged both by tasks that require introspections about self and by tasks that require inferences about the minds of others perceived to be similar to self. To test whether people automatically refer to their own mental states when considering those of a similar other, we examined repetition-related suppression of vMPFC response during self-reflections that followed either an initial reflection about self or a judgment of another person. Consistent with the hypothesis that perceivers spontaneously engage in self-referential processing when mentalizing about particular individuals, vMPFC response was suppressed when self-reflections followed either an initial reflection about self or a judgment of a similar, but not a dissimilar, other. These results suggest that thinking about the mind of another person may rely importantly on reference to one's own mental characteristics.

functional neuroimaging | mentalizing | self-reference | social cognition

Humans consistently explain the behavior of those around them by appealing to others' mental states; that is, their thoughts and feelings, likes and dislikes, current goals and intentions, and enduring dispositions and personality traits (1). Although this understanding of others depends critically on a capacity for rapidly inferring the internal states of those around us (2–4), little is known about how exactly one successfully gains insight into the inner workings of another's mind. After all, no one has ever directly observed the thoughts or feelings of another individual, yet we routinely infer such mental content quickly and easily (5).

One possible solution to the problem of mentalizing may be found in the use of one's own thoughts and feelings as a basis for understanding those of others (6–9). Although the mental states of other people are inherently imperceptible, perceivers do enjoy immediate access to a highly similar system: their own minds. As such, one may infer another person's internal states by spontaneously imagining one's own thoughts, feelings, or desires under similar circumstances and then assuming that the other person would experience comparable mental states, a view alternately described as ''simulationist,'' ''projectionist,'' or ''self-referential'' accounts of social cognition.

Importantly, introspection can only provide insight about another's feelings, beliefs, and preferences to the extent that one's own mind serves as a reasonable proxy for that of the other person. If two people tend to experience very different mental states in the same situations, neither would be well advised to attempt to mentalize about the other on the basis of her own introspection. Thus, the strategy of using one's own mental states as a basis for understanding those of others should be limited to situations in which one can assume that another person generally thinks and feels similarly to oneself. Perceivers may less readily use their own mental states as a guide to the thoughts and feelings of people perceived to be substantially dissimilar from self.

Recently, researchers have used functional neuroimaging to illuminate a specific link between introspection about self and mentalizing about those people perceived to be similar (10, 11). Across several studies, mentalizing about similar versus dissimilar others has been associated with a distinct division of labor in the medial prefrontal cortex, a region ubiquitously identified in neuroimaging studies of mental state inference (12–14). Specifically, a dorsal aspect of the medial prefrontal cortex has been associated with mentalizing about people perceived to be dissimilar from oneself, whereas a more ventral aspect of medial prefrontal cortex (vMPFC) has been linked to mentalizing about those perceived to be similar. Critically, this vMPFC region also has been observed repeatedly during tasks that require participants to introspect about their own mental experiences (15–18), suggesting a connection between tasks that require self-referential thought and those that require inferences about the mental states of similar others.

That the same brain region appears to subserve introspection about oneself and mental state inferences about similar others suggests that an overlapping set of cognitive processes carries out these two otherwise disparate tasks and is consistent with suggestions that perceivers may spontaneously refer to their own mental states to infer those of other people. However, although colocalization of function provides positive evidence that two tasks draw on the same set of mental operations, the limited spatial resolution of hemodynamic imaging techniques, such as fMRI, prevents researchers from using shared functional neuroanatomy as the basis for strong conclusions about the overlap of cognitive process. Because such techniques integrate neural activity across hundreds of thousands of neurons, activation of the same brain voxel by different tasks might occur because each activates distinct, but neighboring or interdigitated, neuronal populations. In this way, two tasks could possibly coactivate the same brain voxel despite engaging different sets of neurons that subserve disparate cognitive processes.

Fortunately, such technical limits can now be circumvented by recently developed paradigms that support stronger conclusions regarding the coactivation of the same neurons by different stimuli or different tasks. These techniques rely on an effect known as ''repetition suppression,'' the observation that neural activity in stimulus-sensitive brain regions is typically reduced when a stimulus is repeated (19). Repetition suppression was initially reported during single-cell recordings in monkeys (20–

Author contributions: A.C.J. and J.P.M. designed research; A.C.J. and J.P.M. performed research; A.C.J. and J.P.M. analyzed data; and A.C.J., C.N.M., and J.P.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

†To whom correspondence should be addressed. E-mail: jenkins@fas.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0708785105/DC1.

22) and has since been observed consistently across a number of studies that measured the fMRI BOLD response in humans (23–29), where it has been used to characterize the response profiles of brain regions involved in a variety of cognitive processes, such as those subserving visual processing, memory, semantics, syntax, number, and motor execution (for reviews, see refs. 19, 30–33).

Although the precise physiological basis of repetition suppression has yet to be fully elucidated, researchers generally agree that the suppressed fMRI BOLD response to repeated stimuli must reflect changes in the firing properties of neurons that subserve the processing of a stimulus, and that suppression across two stimuli indicates that the same (or at least a largely overlapping) population of neurons is engaged by both stimuli (19, 30–33). For example, a demonstration of repetition suppression for the number "3" when it follows "4" but not when it follows "40" might suggest that a relatively high proportion of the neurons that code for the number "3" also participate in representations of similar numerosities (such as "4"), but not in representations of more distant numerosities.

These characteristics of repetition suppression render it well suited for examining the hypothesis that mentalizing about like-minded individuals draws on the same cognitive processes as introspecting about one's own mental characteristics. If (*i*) repeatedly considering one's own mental states produces repetition suppression in self-sensitive regions such as vMPFC, and (*ii*) one engages in self-referential processing when considering the minds of similar others, then (*iii*) repetition suppression also should be observed when perceivers first mentalize about a similar other and then introspect about self. To test these hypotheses, participants in the current study underwent fMRI scanning while answering a series of questions that required introspection about their opinions or preferences (e.g., "How frustrated do you get sitting in traffic?"; see *Methods*). Immediately before each of these self-reflections, participants performed one of three different types of judgments: (*i*) an initial self-reflection; (*ii*) a judgment of the opinions/preferences of a person manipulated to be perceived as similar to self; or (*iii*) a judgment of the opinions/preferences of a person manipulated to be dissimilar from self (participants considered the identical opinion question across phases on half the trials and two different opinion questions for across phases on the other half of trials). Of critical interest was the vMPFC response during self-reflection as a function of the target of the immediately preceding judgment. We expected to observe substantially reduced activity in this region for self-reflections immediately preceded by a prior self-reflection (self-after-self), that is, a significant suppression of the BOLD response when processing the same stimulus category twice consecutively. More important, to the extent that self-referential processing spontaneously accompanies mentalizing about similar others, we expected similar suppression of vMPFC response during self-reflections preceded by judgments of similar others (self-after-similar). In contrast, because referring to one's own mental states should not be an appropriate strategy for mentalizing about dissimilar others, no suppression should be observed when self-reflections follow judgments of dissimilar others (self-after-dissimilar).

Results

Behavioral Data. Postscanning questionnaires confirmed that participants generally held liberal attitudes and perceived themselves to be more similar to the liberal than to the conservative target. On average, participants reported their sociopolitical attitudes as 3.03 on a 7-point scale (1 = very liberal, 4 = neither liberal nor conservative, and 7 = very conservative). Likewise, participants rated the liberal target to be more similar to self ($M = 4.80$ on a 7-point scale) than the conservative target ($M = 3.00$; $P < 0.02$). Moreover, no participant rated the conservative target to be more similar to self than the liberal target. Accordingly, the liberal target was treated as the similar other and the conservative target as the dissimilar other for subsequent fMRI analyses.

Confirming the appropriateness of these target assignments, participants judged the preferences of similar targets to be more closely in line with their own self-reflections than were the preferences of dissimilar targets. For each opinion question, we calculated the mean absolute difference between the participant's self-reported opinion and (*i*) that of the similar other and (*ii*) that of the dissimilar other. Consistent with the notion that one's own preferences may more strongly inform judgments of those perceived to be similar to oneself, judgments of similar others were significantly closer to one's self-reported opinions than were judgments of dissimilar others [Ms = 0.73 vs. 1.12, respectively; $t\ (12) = 3.25$, $P < 0.01$].

fMRI Data. Regions of interest (ROIs) in vMPFC were identified through two independent analyses. We first examined results from the explicit self-reference task, during which participants judged how well an adjective described either their own personality or that of a familiar, but not personally known, other. Consistent with earlier studies (15–18), a single region was obtained from the random effects contrast of self > other trials located in vMPFC (Fig. 1). The response in this region both to self-singletons ($M = 0.06$) and to similar singletons ($M = -0.05$) was significantly greater than the vMPFC response to dissimilar singletons ($M = -0.11$; both Ps < 0.04). Consistent with extant literature suggesting that the vMPFC responds preferentially during judgments about self, the response to similar singletons was intermediate between dissimilar singletons and self-singletons.

This vMPFC ROI was then interrogated for differences among trials on the opinion-judging task, during which participants reported their own opinions/preferences immediately after an initial judgment of self or one of the two other targets. We expected to observe suppression of activity in this vMPFC region for trials on which participants introspected about self immediately after introspecting a first time (i.e., self-after-self) because on these trials participants would be performing the very same task (i.e., reflecting on their own opinion/preference) twice consecutively. Consistent with this prediction, self-after-self judgments were associated with a robust suppression of activity in vMPFC. Indeed, as displayed in Fig. 1, although vMPFC response to self-reflections was typically greater than baseline when participants self-reflected in isolation (i.e., for self-singleton trials), activity in this region was significantly reduced during self-reflections that followed an initial self report. That is, although this vMPFC ROI responded preferentially during self-reflection, its response was substantially suppressed when participants self-reported their opinions twice in a row. Importantly, no difference was observed in the level of repetition suppression during self-after-self trials as a function of whether participants responded to the identical or a different opinion question on the second phase of the trial [$t\ (12) = 0.50$, $P > 0.62$]. Likewise, identical and different judgment pairs both differed significantly from self-singletons (both Ps < 0.05).

In contrast, we expected to observe robust vMPFC activation for self-reflections that immediately followed judgments of dissimilar others (self-after-dissimilar) because perceivers should not spontaneously engage in self-referential thought when mentalizing about those perceived to be different from self, and thus less initial vMPFC processing should occur before the initial self-reflection. Indeed, self-after-dissimilar judgments were associated with a significant increase in vMPFC response over baseline [$t\ (12) = 4.75$, $P < 0.0005$]. This response also was significantly greater than for self-after-self and self-after-similar trials for both identical and different trials (all P values < 0.008),

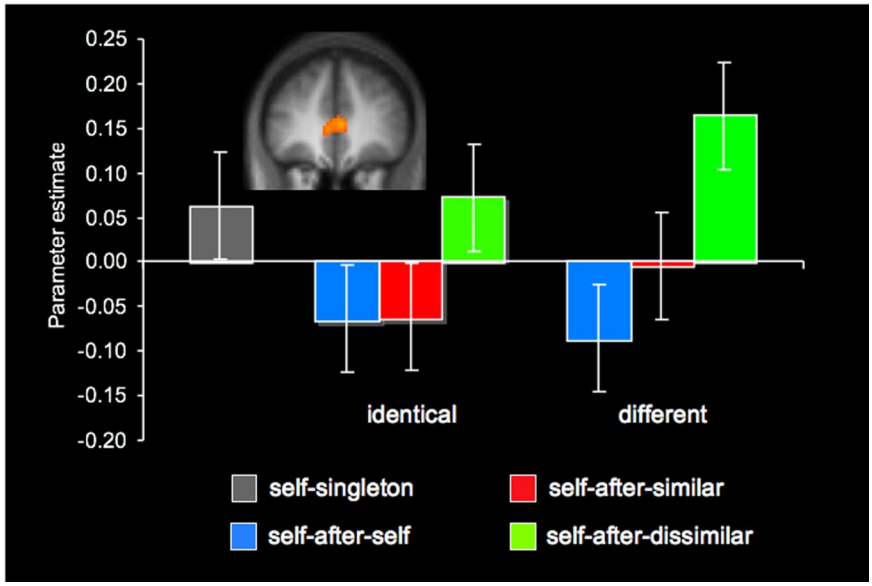

Fig. 1. A region of vMPFC (−6, 45, 3; 47 voxels in extent) was defined from an explicit self-reference task in which judgments of one's own personality characteristics were compared with judgments of another person (i.e., self > other). On a separate task, participants completed a series of paired judgments, in which they introspected about their own preferences and opinions immediately after one of three types of judgments: (*i*) an initial report about self (self-after-self), (*ii*) a judgment of a person with the same sociopolitical attitudes as oneself (self-after-similar), or (*iii*) a judgment of a person with opposing attitudes (self-after-dissimilar). On an equal number of trials, participants considered the identical question for prime and self or a different question across the two phases. The bar graph depicts the BOLD response associated with these self-reports after subtracting out the response associated with the initial judgment (see *Methods*); values therefore represent the additional BOLD response specifically associated with subsequent judgments of self. For comparison purposes, the figure includes the response in this region to self-reports made in isolation (gray bar). Significant repetition suppression was observed for self-reports that followed either an initial self-report (blue bars) or a judgment of a similar other (red bars), but not judgments of a dissimilar other (green bars). Error bars represent the 95% confidence interval for within-subject designs (43).

indicating that the response of vMPFC during self-reflection was not suppressed when participants first mentalized about a dissimilar other.

Of critical interest was whether activity in vMPFC would demonstrate repetition suppression for trials on which participants introspected about self immediately after making a judgment about a person perceived to be similar (self-after-similar). If the same neural processing accompanies both introspection and mentalizing about similar others, the response of vMPFC should fail to distinguish between introspection about self and judgments of similar others. Consistent with this prediction, just as for self-after-self judgments, vMPFC response to self-reflections was substantially suppressed when participants reported their opinion immediately after judging a similar other. For identical judgments, self-after-similar trials were associated with nearly indistinguishable levels of repetition suppression as for self-after-self judgments [$t\ (12) = 0.06, P > 0.95$; for different judgments, $P > 0.11$]. As for self-after-self trials, the response to self-after-similar judgments was negative-going and did not differ significantly from baseline activity (repeated, $P > 0.19$; novel, $P > 0.93$). Together these results suggest that activity in vMPFC, a brain region widely acknowledged to subserve self-referential thought, can be suppressed either by repeatedly introspecting about the self or by introspecting about self immediately after judging a similar, but not a dissimilar, other.

In addition, a second vMPFC ROI was defined from trials within the opinion-judging task by contrasting self-singleton trials to both similar- and dissimilar-singleton trials (i.e., self > other). Random effects analysis identified a region of vMPFC that was preferentially engaged by judgments of self [see supporting information (SI) Fig. 2]. Importantly, the pattern of repetition suppression within this alternative vMPFC ROI was indistinguishable from that in the region defined by the explicit self-reference task (with the exception of self-after-similar-different trials, as detailed below). These results serve to confirm the pattern of findings observed in the vMPFC ROI defined independently by the explicit self-reference task and confirm that vMPFC activity was suppressed for self-reflections that either followed an initial self-reflection or a judgment of a similar, but not dissimilar, other.

Secondary Data and Analyses. In both vMPFC regions, the predicted pattern of repetition suppression was observed across both identical judgments (when participants judged the same opinion question twice within the same trial) and different judgments (when participants judged a different question in the second phase than in the first phase of the trial). The 3 × 2 interaction of pair type (self-after-self, self-after-similar, and self-after-dissimilar) × question repetition (identical or different) did not approach significance in either vMPFC region (both Ps > 0.18), suggesting that the pattern of repetition suppression was similar across identical and different trials. Likewise, when restricted to self-after-similar and self-after-dissimilar trials, the 2 × 2 interaction of trial type × question repetition failed to approach significance in either region (both Ps > 0.55), suggesting that vMPFC activity was similarly suppressed for self-reflections after judgments of similar others, relative to judgments of dissimilar others. Most critically, the same pattern of differences was observed between self-after-similar trials in both regions regardless of whether questions were identical or different. Specifically, vMPFC activity did not differentiate between self-after-similar and self-after-self trials for either identical (both Ps > 0.56) or for different (both Ps > 0.20) trials, but was consistently lower for self-after-similar than self-after-dissimilar (all Ps < 0.05). Finally, pairwise t tests conducted between all trials across repetition (e.g., self-after-self-identical vs. self-after-self-different) revealed a significant difference only for self-after-similar trials in the vMPFC region defined from the explicit

self-reference task ($P < 0.03$) (Fig. 1). Because this effect was not replicated in the alternate vMPFC region identified from self > other from within the judgment task (SI Fig. 2), it remains unclear whether less repetition suppression occurs when a different question is asked about self and a similar other.

Unsurprisingly, participants gave the identical behavioral response for self and other (e.g., responding 3 to both other and self) more often for self-after-similar than self-after-dissimilar targets (37% vs. 29% of trials, respectively; $P < 0.05$). That participants more often made the same behavioral response consecutively for self-after-similar than self-after-dissimilar trials opens the possibility that vMPFC suppression associated with self-after-similar trials could result from repeated motor output, rather than shared cognitive operations between thinking about self and similar others. However, further analysis of the fMRI data belied this possibility. In a secondary analysis of the fMRI data, trials were subconditionalized as a function of whether the same behavioral response was made twice in a row, resulting in four trial types: self-after-similar, same response; self-after-similar, different response; self-after-dissimilar, same response; and self-after-dissimilar, different response (the creation of subconditions was not possible for self-after-self trials, for which prohibitively few different responses were obtained). When analysis was restricted to those trials on which participants made the same behavioral response twice in a row (e.g., pressing 3 for both other and self), we continued to observe repetition suppression for self-after-similar, but not for self-after-dissimilar, judgments, although the difference between the two trial types only reached marginal significance ($P < 0.07$), most likely owing to the reduced power inherent in reducing the number of trials per condition (e.g., as few as 10 trials in a condition for some participants).

In addition, we conducted a secondary analysis restricted only to identical trials, segregating trials on which participants made the same response twice in a row (e.g., 3 to the prime and to self) from those on which participants made two different responses across the two phases of the trial. Critically, the pattern of repetition suppression did not differ as a function of whether participants made the same behavioral response twice in a row to the identical question. First, the difference in the amount of repetition suppression for self-after-similar versus self-after-dissimilar trials did not differ as a function of whether participants made the same behavioral response: the pair type (self-after-similar, self-after-dissimilar) × response overlap (same response, different response) interaction did not approach significance ($F = 1.07$, $P > 0.32$). Second, no simple effect was observed between same versus different responses for either self-after-similar or self-after-dissimilar trials (both Ps > 0.26), suggesting that repetition suppression was not significantly affected by whether participants made the same behavioral response twice in a row. Although these findings are consistent with the interpretation that the pattern of repetition suppression did not differ as a function of making the same behavioral response, results must be interpreted cautiously because these analyses are based on a small subset of the data with reduced power to detect any differences inherent in making the same response or not.

Finally, an accompanying behavioral study reinforced the observation that the facilitation of self-after-similar trials did not result from making a repeated behavioral response. In this data collection, we made use of a behavioral analogue of repetition suppression, wherein repeated processing results in speeded performance on subsequent trials of the same kind (i.e., repetition priming) (33). Specifically, the primary dependent measure in this supporting study was the speed with which a separate group of participants reported their introspection about self after judgments of a similar or dissimilar other. Consistent with the fMRI data, participants ($n = 14$) were significantly faster to self-reflect after a judgment of a similar ($M = 1{,}990$ ms) than a dissimilar ($M = 2{,}079$ ms) target [t (13) = 2.84, $P < 0.02$]. (The parallel analysis of response time in the fMRI experiment was precluded by the abbreviated length of trials necessitated by rapid event-related scanning, such that participants were typically near the ceiling allowed by the response window.) Consistent with the secondary analysis of fMRI data, the significant difference in reaction time between self-after-similar versus self-after-dissimilar was observed even when analysis was restricted to those trials on which participants made the same behavioral response for both self and the other person (M diff = 299 ms; $P < 0.02$). In other words, introspections about one's own attitudes were significantly more facilitated by first mentalizing about a similar than a dissimilar target, even when controlling for participants' tendency to make the same response for self and similar others.

Discussion

These results underscore the tight link between thinking about oneself and thinking about other people, suggesting that self-referential processing may be triggered spontaneously when considering the mental states of others. Using two different sets of contrasts, we identified ROIs in vMPFC that, consistent with earlier studies of self-reference (15–18), responded preferentially during trials that required introspection about one's own mental characteristics. We then examined the pattern of response in these ROIs during an opinion-judging task in which participants reported their preferences and opinions immediately after a prior self-reflection, a judgment of a similar other, or a judgment of a dissimilar other. Extending earlier observations that repeatedly processing the same stimulus leads to the suppression of the activity in task-sensitive brain regions, the response of vMPFC was attenuated for self-reflections that immediately followed a preceding introspection about self. Critically, this same repetition suppression was observed for self-reflections that followed judgments of similar, but not dissimilar, targets. That is, whereas self-reflections were associated with a significant activation of vMPFC after judgments of a dissimilar other, the response of this region was equivalently suppressed during self-reflections that followed judgments of a similar other as during those that followed initial self-reflections, suggesting that vMPFC failed to discriminate between self-referential thought and mentalizing about a similar other.

The use of repetition suppression provides particularly strong support for the conclusion that mentalizing about similar others draws on the same cognitive processes as introspecting about oneself. Most proposals regarding the physiological basis of repetition suppression conclude that the likely basis of the effect is that the same, or largely overlapping, population of neurons subserves the processing of two distinct stimuli. These changes may include neurons "fatiguing" upon repeated firings, briefer neuronal firing durations, or the recruitment of fewer total neurons to process stimuli a second time (19, 33, 34). Regardless of the precise nature of the neuronal change, observing repetition suppression in a brain region across two seemingly distinct stimuli provides evidence that the neurons in that region are insensitive to any difference between the stimuli. This interpretation of repetition suppression suggests the intriguing possibility that the population of vMPFC neurons subserving introspections about self serve double-duty by also participating in representing the minds of similar others. Accordingly, these results contribute additional support for simulationist views of social cognition, which have suggested that one important mechanism for understanding the thoughts and feelings of others is reference to one's own mental states.

The observed dissociation between the functional neuroanatomy associated with mentalizing about similar and dissimilar others joins several other recent observations that likewise

suggest that the cognitive processes deployed during mentalizing will vary as a function of exactly whose mental states one is attempting to infer (10, 11). When another person is assumed to be sufficiently similar to self, perceivers appear to make use of the same processes deployed for introspecting about their own mental characteristics, but decline to do so for others assumed to be dissimilar from self. Such findings reinforce the emerging view that social cognition, rather than being composed of a single, all-purpose module for mentalizing (35), relies on a number of different strategies for mentalizing that vary with the particulars of the social environment (36–39). Of course, exactly how perceivers determine whether a particular individual is sufficiently similar to justify the use of self-referential mentalizing remains an open question, as does characterization of the cognitive processes that subserve mentalizing about dissimilar others.

Putting introspection to use in mentalizing undoubtedly provides a rich starting point for contemplating the minds of others, allowing perceivers to bring to bear the full complement of their own attitudes, feelings, and beliefs in inferring those of another person. Somewhat ironically, however, because one's own mental states may serve as an appropriate proxy only for those whom we assume think and feel like we do, human social cognition may possess an intrinsic bias to discriminate between those perceived to be similar, like-minded members of one's ingroup and those perceived to be dissimilar, exotic others. Indeed, that our minds naturally segregate dissimilar from similar others and then mentalize in a distinct way about those perceived to be different from self may be one of the factors that gives rise to aspects of outgroup prejudice, such as stereotypes about members of various racial, ethnic, or cultural backgrounds. Accordingly, one strategy for successfully counteracting such biases may be to augment the degree to which perceivers engage in self-referential mentalizing about otherwise dissimilar others, for example, by consciously taking the perspective of another person (44). Like many of the cognitive heuristics that typically serve us well, but periodically lead to undesirable or maladaptive behavior (40), the use of self-reference in mentalizing may be a double-edged sword: a useful strategy for providing rich and accurate insights into the minds of similar individuals, but rife with the potential to exclude those minds assumed at first glance to be different from our own.

Methods

Participants. Participants were 13 (8 male) right-handed, native English speakers with no history of neurological problems (mean age 20.7 years, range 19–23). One additional participant, who was being treated for depression at the time of the study, was excluded from analysis. All participants were undergraduate or graduate students at universities in the Boston area, and all provided informed consent in a manner approved by the Human Studies Committee of the Massachusetts General Hospital.

Stimuli and Behavioral Procedure. Participants were told that the experiment investigated the ability to make inferences about others on the basis of minimal information. Before scanning, participants read a short paragraph about each of two unfamiliar target individuals depicted by face photographs. Following Mitchell *et al.* (11), one target was described as a college student in the Northeast who maintained liberal social and political attitudes similar to those of our typical student participant. In contrast, the other target was described as a conservative, fundamentalist Republican attending a large university in the Midwest (i.e., as fairly dissimilar from our typical participant). Targets were always the same sex as the participant, and both the pairing of particular faces to descriptions and the order of presentation (liberal target first, conservative target first) were randomized across participants. Participants were given as much time as needed to read about each of the two targets.

During scanning, participants performed a modified version of the opinion-judging task used by Mitchell *et al.* (11) Trials were divided into prime and self phases. Each trial began with the presentation of one of three primes: (*i*) the photograph of the liberal target, (*ii*) the photograph of the conservative target, or (*iii*) a chalk outline of a head with the word ''me'' written inside, used to represent the participant her or himself. This prime image appeared above a four-point response scale (1 = not at all and 4 = definitely). Simultaneously, an opinion question appeared between the prime and the response scale, and participants were asked to use the scale either to estimate how likely the target would be to endorse the opinion or, for the chalk outline, to report their own response to the question. Opinion questions referred to a range of personal issues that were pretested to be unrelated to political orientation (e.g., ''dislike mushrooms on pizza?''; ''enjoy crossword puzzles?''; ''like to be the center of attention?''; ''generally see things from many perspectives?''; ''enjoy helping friends with problems?''; and ''like impressionist artwork?''). The prime image, question, and scale remained onscreen together for 3,600 ms.

The self phase of each trial began after a 400-ms interval and was identical to the chalk outline prime described above, in which participants reported their own response to an opinion question. Self-judgments were conditionalized as a function of the preceding target, resulting in three trial types: self-after-similar, self-after-dissimilar, and self-after-self. Because participants were identified as having liberal sociopolitical attitudes, the liberal target was designated ''similar'' and the conservative target was designated ''dissimilar''; the validity of these target assignments was confirmed by postscan questionnaires that asked participants to rate how similar they perceived each of the two targets to be relative to self.

For half the trials, the same opinion question was asked in both the prime and self phases of the trial (identical trials). In the remaining half of the trials, a different question was asked in the two phases (different trials). However, participants were instructed to consider each question individually. Although none of the primary analyses was qualified by whether the identical or different question was asked across the two phases, we report data separately for these two trial types for the sake of completeness.

Participants completed 240 such paired prime–self trials. In addition, the experimental design included 90 singleton trials, on which participants saw only the prime phase (30 each of self, similar, and dissimilar) without a subsequent self phase. These singletons were included as catch trials used to facilitate deconvolution of the hemodynamic response specifically associated with the self phase (see below).

After the opinion-judging task, participants completed an explicit self-reference task that has been used previously to identify a region of vMPFC that responds preferentially during self-referential judgments (15, 16). On each of 100 trials, participants saw a single trait adjective that could be used to describe a person's personality or dispositional traits (e.g., curious, intelligent, or neurotic). Each trait adjective was accompanied by the name of one of two targets: self or Bush. For self-trials, participants were asked to use a 4-point scale to indicate how well the trait adjective described themselves. For Bush trials, participants were asked to use the scale to indicate how well the adjective described the current U.S. president, George W. Bush. This choice of other was guided by earlier studies of self-referential processing, which have typically used the current head of state (a familiar, but not personally known, other) as a comparison to self-judgments (15, 16, 41). To optimize estimation of the event-related fMRI response, on both the opinion-judging and explicit self-reference tasks, trials were intermixed in a pseudorandom order and separated by a variable interstimulus interval (400–8,000 ms) (42), during which participants passively viewed a fixation crosshair.

After scanning, participants answered two questions about their own sociopolitical attitudes in random order (''How politically liberal or conservative are you?'' and ''How socially liberal or conservative are you?'') by using a 7-point scale (1 = very liberal, 4 = neither liberal nor conservative, and 7 = very conservative). Finally, participants reported how similar they perceived each of the two targets to be to themselves (1 = most dissimilar to 7 = most similar).

Imaging Procedure. FMRI data were collected by using a 3 Tesla Siemens Trio scanner. The opinion-judging task comprised five functional runs of 296 volume acquisitions, and the explicit self-reference task comprised two functional runs of 130 volume acquisitions (26 axial slices, 5 mm thick; 1 mm skip). Functional imaging used a gradient-echo echo-planar pulse sequence (TR = 2 s; TE = 35 ms; 3.75 × 3.75 in-plane resolution). After the functional scans, we collected a high-resolution T1-weighted structural scan (MP-RAGE). PsyScope software for Mac OS X (L. Bonatti, International School of Advanced Studies, Trieste, Italy) was used to project stimuli onto a screen at the end of the magnet bore, which participants viewed via a mirror mounted on the head coil. A pillow and foam cushions were placed inside the coil to minimize head movement.

FMRI data were preprocessed and analyzed by using SPM99 (Wellcome Department of Cognitive Neurology, London, United Kingdom). 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)] by using a Gaussian kernel.

Statistical analyses were performed by using the general linear model in which the event-related design was modeled by 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 into a second-level analysis, treating participants as a random effect.

First, a region of vMPFC was identified from the comparison of self > other (i.e., self trials vs. Bush trials). In addition, a second vMPFC ROI was defined from within the opinion-judging task from the comparison of self > other by using singleton trials only (i.e., self-singletons > similar singletons plus dissimilar singletons). Peak coordinates were identified by using a statistical criterion of 25 or more contiguous voxels at a voxel-wise threshold of $P < 0.001$. This cluster size was selected on the basis of a Monte Carlo simulation (S. Slotnick, Boston College, Boston) of our brain volume, which indicated that this cluster extent cutoff provided an experiment-wise threshold of $P < 0.05$, corrected for multiple comparisons.

For the opinion-judging task, trials were conditionalized as a function of (*i*) whether they were paired or singleton, and (*ii*) the identity of the prime, resulting in six trial types: self-after-self, self-after-similar (i.e., judgment of the liberal target, then self), self-after-dissimilar, self-singleton, similar singleton, and dissimilar singleton. The parameter estimates associated with each of these six trial types were extracted from the two vMPFC ROIs identified after the above procedure. Of critical interest was the extent to which brain activity associated with the self phase was suppressed as a function of the identity of the preceding prime (self, similar, and dissimilar). Because the self phase was always preceded by a prime (i.e., self and prime were intrinsically confounded), we obtained such a measure of repetition suppression by subtracting the response to singleton trials from the corresponding paired prime–self trials. For example, the response that was associated with the self phase of self-after-similar trials was indexed as the difference of self-after-similar trials minus similar-singleton trials. The resulting scores represent activity that is specifically associated with the self portion of trials as a function of whether this judgment was made immediately after an initial judgment of self, of a similar other, or a dissimilar other.

ACKNOWLEDGMENTS. We thank D. L. Ames, M. Banaji, R. Buckner, Y. Jiang, and L. Powell for advice and assistance. This work was supported by National Science Foundation Grant BCS 0642448 (to J.P.M.), the Athinoula A. Martinos Center for Biomedical Imaging, National Center for Research Resources Grant P41RR14075, the Mental Illness and Neuroscience Discovery Institute, and a Royal Society-Wolfson Fellowship (to C.N.M.).

1. Dennett DC (1987) *The Intentional Stance* (MIT Press, Cambridge, MA).
2. Tomasello M (1999) *The Cultural Origins of Human Cognition* (Harvard Univ Press, Cambridge, MA).
3. Gilbert DT (1998) in *Handbook of Social Psychology*, eds Gilbert DT, Fiske ST, Lindzey G (McGraw–Hill, New York), pp 89–150.
4. Frith CD, Frith U (1999) *Science* 286:1692–1695.
5. Apperly IA, Riggs KJ, Simpson A, Chiavarino C, Samson D (2006) *Psychol Sci* 17:841–844.
6. Gordon RM (1992) *Mind and Language* 1:158–171.
7. Heal J (1986) in *Language, Mind and Logic*, ed Butterfield J (Cambridge Univ Press, Cambridge, UK), pp 135–150.
8. Davies M, Stone T, eds (1995) *Mental Simulation: Evaluations and Applications* (Blackwell, Oxford).
9. Nickerson R (1999) *Psychol Bull* 125:737–759.
10. Mitchell JP, Banaji MR, Macrae CN (2005) *J Cogn Neurosci* 17:1306–1315.
11. Mitchell JP, Macrae CN, Banaji MR (2006) *Neuron* 50:655–663.
12. Blakemore SJ, Winston J, Frith U (2004) *Trends Cogn Sci* 8:216–222.
13. Frith C, Frith U (2001) *Curr Dir Psychol Sci* 10:151–155.
14. Mitchell JP (2006) *Brain Res* 1079:66–75.
15. Kelley WM, Macrae CN, Wyland CL, Caglar S, Inati S, Heatherton TF (2002) *J Cogn Neurosci* 14:785–794.
16. Macrae CN, Moran JM, Heatherton TF, Banfield JF, Kelley WM (2004) *Cereb Cortex* 14:647–654.
17. Zysset S, Huber O, Ferstl E, von Cramon DY (2002) *NeuroImage* 15:983–991.
18. Johnson SC, Baxter LC, Wilder LS, Pipe JG, Heiserman JE, Prigatano GP (2002) *Brain* 125:1808–1814.
19. Grill-Spector K, Henson R, Martin A (2006) *Trends Cogn Sci* 10:14–23.
20. Li L, Miller EK, Desimone R (1993) *J Neurophysiol* 69:1918–1929.
21. Miller EK, Desimone R (1994) *Science* 263:520–522.
22. Sobotka S, Ringo JL (1996) *J Neurosci* 16:4222–4230.
23. Demb JB, Desmond JE, Wagner AD, Vaidya CJ, Glover GH, Gabrieli JD (1995) *J Neurosci* 15:5870–5878.
24. Stern CE, Corkin S, Gonzalez RG, Guimaraes AR, Baker JR, Jennings PJ, Carr CA, Sugiura RM, Vedantham V, Rosen BR (1996) *Proc Natl Acad Sci USA* 93:8660–8665.
25. Henson R, Shallice T, Dolan R (2000) *Science* 287:1269–1272.
26. Buckner RL, Petersen SE, Ojemann JG, Miezin FM, Squire LR, Raichle ME (1995) *J Neurosci* 15:12–29.
27. Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, Malach R (1999) *Neuron* 24:187–203.
28. Jiang Y, Haxby JV, Martin A, Ungerleider LG, Parasuraman R (2000) *Science* 287:643–646.
29. Vuilleumier P, Henson RN, Driver J, Dolan RJ (2002) *Nat Neuro* 5:491–499.
30. Grafton ST, Hamilton AFdC (2007) *Hum Mov Sci* 26:590–616.
31. Grill-Spector K, Malach R (2001) *Acta Psychol* (*Amst*) 107:293–321.
32. Sawamura H, Orban GA, Vogels R (2006) *Neuron* 49:307–318.
33. Wiggs CL, Martin A (1998) *Curr Opin Neurobiol* 8:227–233.
34. Henson RN, Goshen-Gottstein Y, Ganel T, Otten LJ, Quayle A, Rugg MD (2003) *Cereb Cortex* 13:793–805.
35. Leslie AM, Friedman O, German TP (2004) *Trends Cogn Sci* 8:528–533.
36. Ames DR (2004) *J Pers Soc Psychol* 87:340–353.
37. Ames DR (2004) *J Pers Soc Psychol* 87:573–585.
38. Macrae CN, Bodenhausen GV, Milne AB (1994) *J Pers Soc Psychol* 66:37–47.
39. Malle BF (2005) in *Other Minds: How Humans Bridge The Divide Between Self and Other*, eds Malle BF, Hodges SD (Guilford, New York), pp 26–43.
40. Tversky A, Kahneman D (1974) *Science* 27:1124–1131.
41. Rogers TB, Kuiper NA, Kirker WS (1977) *J Pers Soc Psychol* 35:677–688.
42. Dale AM (1999) *Hum Brain Mapp* 8:109–114.
43. Loftus GR, Masson MEJ (1994) *Psychonomic Bull Rev* 1:476–490.
44. Ames DL, Jenkins AC, Banaji MR, Mitchell JP (2008) *Psychol Sci*, in press.

Cerebral Cortex February 2010;20:404–410
doi:10.1093/cercor/bhp109
Advance Access publication May 28, 2009

Mentalizing under Uncertainty: Dissociated Neural Responses to Ambiguous and Unambiguous Mental State Inferences

Adrianna C. Jenkins and Jason P. Mitchell

Department of Psychology, Harvard University, Cambridge, MA 02138, USA

The ability to read the minds of others (i.e., to mentalize) requires that perceivers understand a wide range of different kinds of mental states, including not only others' beliefs and knowledge but also their feelings, desires, and preferences. Moreover, although such inferences may occasionally rely on observable features of a situation, perceivers more typically mentalize under conditions of "uncertainty," in which they must generate plausible hypotheses about a target's mental state from ambiguous or otherwise underspecified information. Here, we use functional neuroimaging to dissociate the neural bases of these 2 distinct social–cognitive challenges: 1) mentalizing about different types of mental states (beliefs vs. preferences) and 2) mentalizing under conditions of varying ambiguity. Although these 2 aspects of mentalizing have typically been confounded in earlier research, we observed a double dissociation between the brain regions sensitive to type of mental state and ambiguity. Whereas ventral and dorsal aspects of medial prefrontal cortex responded more during ambiguous than unambiguous inferences regardless of the type of mental state, the right temporoparietal junction was sensitive to the distinction between beliefs and preferences irrespective of certainty. These results underscore the emerging consensus that, rather than comprising a single mental operation, social cognition makes flexible use of different processes as a function of the particular demands of the social context.

Keywords: medial prefrontal cortex, mentalizing, neuroimaging, social cognition, theory of mind

Introduction

Unlike encounters with falling tree branches, stalled cars, or other inanimate objects, an understanding of other people requires the tacit recognition that their behavior is influenced by the contents of their minds (Dennett 1987). However, the ability to infer the nature of those contents—that is, to mentalize—poses a series of nontrivial challenges to human cognition. Perceivers only rarely receive explicit reports about another person's thoughts, feelings, or desires and must instead interpret ambiguous hints about the hidden inner workings of other minds: for example, attempting to uncover the possible significance of an eyebrow raise, sidelong glance, vocal inflection, or sudden departure. Each of these bits of information, in turn, may be clues to a wide range of possible kinds of mental states, such as what a person is thinking (i.e., beliefs), feeling (emotions), desiring (wants and preferences), or intending (goals). Finally, having generated a provisional model of another person's mind, perceivers must also calculate how the contents of that mind are likely to influence the person's behavior.

Given the complexity and diversity of the inferences we make about others, humans likely developed a suite of cognitive processes that, together, allow us to traffic so readily in the mental worlds of other people. Consistent with this possibility that social cognition comprises several distinct processes that meet different computational demands, researchers have identified a set of several brain regions that respond consistently when considering the minds of others: dorsal and ventral aspects of the medial prefrontal cortex (MPFC), the temporoparietal junction (TPJ), medial parietal cortex, and the superior temporal sulcus (Fletcher et al. 1995; Goel et al. 1995; Gallagher et al. 2000, 2002; Mitchell et al. 2002; Saxe and Kanwisher 2003; Van Overwalle 2009). Having identified this constellation of regions involved in human social abilities, researchers have now begun to isolate specific mental processes subserved by each, with the aim of decomposing social cognition into its constituent parts.

Importantly, the main challenge in this enterprise has been delineating the dimensions along which social cognition might be expected to divide. One natural starting place has been the observation that perceivers must infer a variety of different types of mental states, such as beliefs, feelings, and intentions, and indeed, researchers have recently suggested that different brain regions may subserve mentalizing about these different kinds of mental content. For example, a right-lateralized region of TPJ has been implicated specifically in representing others' beliefs (Saxe and Kanwisher 2003; Saxe and Powell 2006), and MPFC has emerged consistently from tasks involving inferences about affective states or preferences (Mitchell, Banaji, Macrae 2005a; Hynes et al. 2006; Mitchell et al. 2006; Vollm et al. 2006; Shamay-Tsoory and Aharon-Peretz 2007). Taken together, these observations have led some commentators to conclude that activation in TPJ and MPFC may be modulated specifically by differences among particular types of mental content to be inferred (e.g., Van Overwalle 2009).

However, in addition to inferring different types of mental states, humans must also mentalize under varying degrees of certainty. In some situations, an inference about the state of another person's mind is all but dictated by given information. For example, when Sarah puts her cookie in the office refrigerator and returns to retrieve it 5 min later, we are fairly confident that she "believes" her cookie is in the refrigerator. Similarly, if Sarah always chooses oatmeal cookies from her many dessert options, we can be fairly confident that Sarah "likes" or "prefers" oatmeal cookies. In these cases, perceivers' inferences can be formulated using a simple set of rules operating over explicit, observable information about a target. To infer where Sarah thinks her cookie will be 5 min after she stashes it in the refrigerator, perceivers may simply apply the rule that people generally can recall easily what

they did 5 min ago. Likewise, perceivers may conclude that Sarah has a particular fondness for oatmeal cookies by applying the rule that if someone freely and consistently chooses an object (e.g., oatmeal cookies) over comparable alternatives, then that person likely prefers that object (Kelley 1972). In both cases, readily observable information can feed into some basic social–cognitive rules to produce fairly unambiguous inferences about another person's mental states.

In contrast, many inferences about human minds take place under conditions of far greater ambiguity. When Steve arrives home and hears voices inside his apartment, will he believe that he is being robbed, that he accidentally left the TV on, that his parents have made a surprise visit, or something else? Similarly, if Steve always arrives late to lecture when the only available seats are in the back of the room, we cannot be particularly confident that he really does prefer to sit far away from the professor. Because the information in such situations is insufficient to constrain one's inferences fully, perceivers must make do with provisional hypotheses about a target's mental states, which remain ambiguous until further clues about their contents are discerned. Although perceivers do make assumptions about other minds even under conditions of relative ambiguity (Gilbert 1998), it is unlikely that they do so using the kind of rule-based processes that can be brought to bear more fruitfully for inferences of greater certainty. Rather, given a scarcity of suitably definitive inputs to our social–cognitive rules, mentalizing under uncertainty likely relies on an alternative, more flexible, and internally generated system for making sense of other minds.

In attempting to identify the dimensions along which social cognition dissociates, most extant research has confounded differences in mentalizing about varying types of internal states with differences in mentalizing under varying degrees of certainty. For example, although the TPJ has been specifically linked to a particular type of mental state—beliefs—the information provided in typical belief mentalizing tasks essentially dictates the mental state of the protagonist, making perceivers' inferences unambiguous. In the bulk of experiments identifying the TPJ with beliefs (Saxe and Kanwisher 2003; Samson et al. 2004; Saxe and Wexler 2005; Saxe and Powell 2006), perceivers read stories based on the classic "Sally–Anne" problem developed for use in children: perceivers watch Sally place her ball in a basket and then, while Sally is away and unaware, they watch Anne surreptitiously move the ball to a second location, at which point they are asked where Sally will look for her ball when she returns. This kind of situation contains all the information needed for an unambiguous, rule-based inference about what Sally believes or thinks (i.e., that the ball is still safely hidden in its original location). In contrast, the information provided in typical preference or affective mentalizing tasks leaves inferences much more open-ended (Hynes et al. 2006; Mitchell et al. 2006; Vollm et al. 2006; Shamay-Tsoory and Aharon-Peretz 2007). For example, participants might be told that Sarah is politically liberal and subsequently be asked whether she would prefer to go hiking or go to the beach (Mitchell et al. 2006). The frequent conflation of these 2 dimensions raises the possibility that findings previously attributed to differences in type of mental state, such as the preferential engagement of MPFC during inferences about others' preferences, may in fact be better attributed to differences in the certainty with which such mental state inferences can be made.

Indeed, a substantial amount of other research supports the possibility that MPFC may subserve mentalizing under uncertainty rather than inferences about particular types of mental states per se. Recently, this region has been implicated in processes supporting the ability to draw on elements of relevant past experiences in order to formulate novel predictions (Addis et al. 2007; Buckner and Carroll 2007), as well as in the use of one's own experience to mentalize about others (Mitchell et al. 2006; Jenkins et al. 2008). When inferences are relatively underspecified by situational constraints, perceivers may find it especially useful to mentalize on the basis of such simulated, internally generated information, whether that information arises from their own firsthand experience or from having observed similar circumstances in the past. That is, perceivers may find it particularly useful to rely on associations formed through past experiences as they generate predictions about what another person may be thinking or feeling in ambiguous or uncertain situations (Mitchell forthcoming). In contrast, such a process may be less useful under circumstances in which another person's mental state could be inferred simply by applying general "rules" about human minds.

In the current experiment, we investigated the extent to which regions associated with mentalizing would be modulated independently by type of mental state (beliefs vs. preferences) and the uncertainty surrounding one's inference about it. Participants were scanned using functional magnetic resonance imaging (fMRI) as they read short vignettes that supported either an unambiguous or ambiguous inference about a person's beliefs or preferences. Unambiguous versions of each vignette were written such that the information in the scenario would strongly suggest the mental state of the protagonist, whereas ambiguous vignettes implied that the protagonist's mental state could be any one of multiple possibilities. For each vignette, participants were obliged to consider either the protagonist's beliefs or his or her preferences, thus allowing us to dissociate brain regions that were sensitive to differences in mental state type (belief and preference) from those sensitive to differences in mentalizing certainty (ambiguous and unambiguous). Although interested in the potential effects of these dimensions across the brain, we were particularly interested in examining the extent to which MPFC contributions to social cognition are better characterized as subserving inferences about affectively laden mental states (such as preferences) or as more generally subserving ambiguous inferences under uncertainty. Moreover, this design also allowed us to test earlier claims that the TPJ specifically subserves inferences about a particular type of mental state (i.e., beliefs).

Materials and Methods

Participants

Fifteen right-handed college undergraduates (9 females, age range 18–22 years, mean age 19.8 years) with no history of neurological problems participated in exchange for pay or course credit. Participants provided informed consent in accordance with the guidelines maintained by Massachusetts General Hospital.

Stimuli and Behavioral Procedure

Mentalizing Task

During scanning, participants read short vignettes relating the events of an everyday scenario. Vignettes conveyed information about either

a protagonist's beliefs or preferences (see Table 1 for examples and Supplementary Material for full stimulus set). Unambiguous versions of each scenario were written such that the information in the scenario would strongly suggest, but not state explicitly, the belief or preference of the protagonist. Such scenarios relied heavily on perceptual truisms about human beings (e.g., that they generally perceive objects in the environment and generally remember what they have recently seen). Ambiguous versions of each scenario were written such that the protagonist's belief or preference could plausibly be any one of multiple possibilities under the circumstances provided, that is, the information given did not dictate a correct response but rather left the inference more open-ended. A slight change in what would otherwise be an unambiguous scenario might render deterministic rules about the human mind inapplicable and the scenario therefore ambiguous: for example, if Sarah, on her way out the door having just put her cookie in the refrigerator, hears Tom tell her he's moving her cookie but he does not say where (belief) or if Sarah always eats an oatmeal cookie after dinner but there are never any other options because Tom always buys dessert (preference), we can be less certain about Sarah's mental states given the information provided. Unambiguous belief vignettes were created in both "true belief" and "false belief" versions; however, no differences were observed between true and false beliefs, and analyses were therefore collapsed across this dimension.

Stimuli in all 4 groups (ambiguous preference, ambiguous belief, unambiguous preference, and unambiguous belief) were matched for length (mean number of characters = 213.5). Matched control stories in which participants inferred the content of physical representations (such as those in photographs or on maps) were used for comparison (Zaitchik 1990). For example, participants might read about a tree house that was photographed before being painted blue and be asked to identify the color in which it would have appeared in the photo (Table 1).

Following each mentalizing scenario, participants answered a single multiple-choice question about the protagonist's belief or preference; following each nonsocial scenario, participants answered a question about a physical representation (such as a map or photograph). For all scenarios, the story and question remained onscreen together for a total of 10 s, at which point the story disappeared and 4 response choices were presented for 4 s. In all conditions, participants were asked to formulate an answer to every question before any response choices appeared. Accordingly, to allow for the possibility that participants generated ideas other than those represented by our answer choices, the fourth response option was always a none-of-the-above possibility (e.g., "Somewhere else"). Each trial was followed by 12 s of fixation. Each participant completed a total of 60 mentalizing scenarios and 12 nonsocial scenarios across 4 functional runs, with presentation randomized across participants such that no participant ever encountered both an ambiguous and an unambiguous version of the same story.

Imaging Procedure

fMRI data were collected using a 3-T Siemens Trio scanner across 4 functional runs of 234 volume acquisitions (26 axial slices, 5 mm thick, 1 mm skip). Functional imaging used a gradient-echo echo planar pulse sequence (time repetition = 2 s, time echo = 35 ms, 3.75 × 3.75 in-plane resolution). Prior to the functional scans, we collected a high-resolution T_1-weighted structural scan (magnetization-prepared rapid gradient echo). PsyScope software for Mac OS X (L. Bonatti, International School of Advanced Studies, Trieste, Italy) was used to project stimuli onto a screen at the end of the magnet bore, which participants viewed via a mirror mounted on the head coil. A pillow and foam cushions were placed inside the coil to minimize head movement.

fMRI data were preprocessed and analyzed using SPM2 (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) using a Gaussian kernel.

Statistical analyses were performed using the general linear model in which the blocked design was modeled using a boxcar function 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 into a second-level analysis treating participants as random effect. Peak coordinates were identified using a statistical criterion of 25 or more contiguous voxels at a voxelwise threshold of $P < 0.0001$. Monte Carlo simulations (S. Slotnick, Boston College) of our brain volume confirmed that these criteria provided a brainwise alpha level of $P < 0.05$, corrected for multiple comparisons.

Table 1
Stimulus examples

	Unambiguous	Ambiguous
Belief	Pam is an avid gardener. The weather was so warm today that all the tulips in Pam's backyard suddenly bloomed. The tulips next to Pam's office still have not yet flowered, though. Pam has been at work all day.	Pam is an avid gardener and is particularly fond of her tulips. It's early spring, and a few of her flowers have begun to bloom. When Pam got home from work today, her neighbor told her she might want to take a look at her tulip beds.
	What does Pam think?	What does Pam think?
	1. Her tulips have bloomed	1. Her tulips have bloomed
	2. Her tulips have not bloomed yet	2. Her tulips have not bloomed yet
	3. Her tulips have died	3. Her tulips have died
	4. Something else	4. Something else
Preference	Erin has 2 classes on Tuesdays. Today was the last day of Tuesday classes. In both of her classes, Erin is usually one of the first people there, and she always sits in the back.	Erin has 2 classes on Tuesdays. Today was the first day of Tuesday classes for the semester. In both of her classes, the room was quite full when Erin arrived, and she sat in the back.
	Where does Erin like to sit in class?	Where does Erin like to sit in class?
	1. In the front	1. In the front
	2. In the back	2. In the back
	3. In the middle	3. In the middle
	4. Somewhere else	4. Somewhere else
Nonsocial	The color printer cartridge just ran out of blue ink, but it kept printing anyway. It printed a picture of a healthy grass lawn from a computer screen.	
	In the printed picture, what color is the grass?	
	1. Yellow	
	2. Green	
	3. Blue	
	4. Something else	

Note. Mentalizing scenarios support inferences that differ the type of mental state to be inferred (belief vs. preference) and the certainty with which the inference can be made (unambiguous vs. ambiguous). Nonsocial scenarios support inferences without mental content.

We first identified regions of interest from the comparison of mentalizing > nonsocial (i.e., all unambiguous and ambiguous belief and preference stories vs. nonsocial control stories). These regions were then interrogated for differences among the mentalizing scenarios by comparing the parameter estimates associated with the 4 mentalizing trial types: unambiguous belief, unambiguous preference, ambiguous belief, and ambiguous preference. To confirm the results of the region-of-interest analysis, we also conducted whole-brain, random-effects analyses of unambiguous versus ambiguous scenarios (collapsing across content type) and belief versus preference scenarios (collapsing across ambiguity).

Results

Behavioral Results

To confirm our conditionalization of scenarios as unambiguous and ambiguous, we first examined the distribution of participants' responses to each question as a function of ambiguity. Specifically, for each question, we calculated the proportion of participants who chose the most commonly selected of the 3 possible contentful answers, excluding "none of the above" responses. Participants overwhelmingly converged on a single answer to each unambiguous scenario, choosing the modal response 87% of the time. In contrast, responses were more variable for ambiguous scenarios, with participants agreeing on a single answer only 51% of the time, χ^2 (1, n = 15) = 8.14, P < 0.005). Moreover, for unambiguous questions, participants chose "none of the above" less than 1% of the time, whereas for ambiguous questions, participants made use of this option 30% of the time, χ^2 (1, n = 15) = 27.13, P < 0.0001. This pattern of responding was observed for both beliefs (88% agreement for unambiguous belief vs. 53% agreement for ambiguous belief inferences) and preferences (84% agreement for unambiguous preference vs. 51% agreement for ambiguous preference inferences).

fMRI Results

The primary question of interest was the extent to which brain regions involved in mentalizing would be sensitive to differences in content type and in ambiguity. To identify brain regions involved in mentalizing, we first conducted a whole-brain, random-effects analysis of all "mentalizing > nonsocial" scenarios. This contrast revealed a set of regions commonly associated with social cognition, including both dorsal and ventral MPFC (vMPFC), right and left TPJ, the superior temporal sulcus, the temporal poles, and medial parietal cortex (Fig. 1). We then interrogated these regions of interest for their sensitivity to content type, ambiguity, and the interaction between these 2 factors. The response of 3 regions was modulated by ambiguity and/or content (see Table 2). First, dorsal MPFC (dMPFC) demonstrated greater response during ambiguous than unambiguous inferences, $F_{1,14}$ = 7.44, P < 0.02, d = 0.73, but did not differentiate between preferences and beliefs, $F_{1,14}$ = 1.72, P > 0.21, d = 0.35. In contrast, right TPJ was characterized by greater activation during belief than preference scenarios, $F_{1,14}$ = 9.74, P < 0.01, d = 0.83, but did not differentiate between scenarios as a function of ambiguity, $F_{1,14}$ = 0.79, P > 0.38, d = 0.24, consistent with suggestions that right TPJ contributes specifically to mentalizing about beliefs (Saxe and Kanwisher 2003; Saxe and Powell 2006). Finally, activation in vMPFC was characterized by main effects of both ambiguity, $F_{1,14}$ = 7.10, P < 0.02, d = 0.71, and content, $F_{1,14}$ = 12.26, P < 0.0005, d = 0.94, such that the region responded more during ambiguous than unambiguous inferences and also responded more during inferences about preferences than during inferences about beliefs. All 3 regions showed no evidence of an ambiguity × content interaction, all F values < 1.65, Pvalues > 0.22. In contrast, a marginally significant interaction between type and ambiguity was observed in medial parietal cortex, $F_{1,14}$ = 3.03, P < 0.10, d = 0.47, such that the region responded more during unambiguous than ambiguous inferences about beliefs but more during ambiguous than unambiguous inferences about preferences; however, neither the main effect of ambiguity (P > 0.45) nor the main effect of content (P > 0.76) approached significance in this region. Moreover, the presence of a significant 2-way interaction of region × content, $F_{2,42}$ = 9.10, P < 0.001, confirmed that the pattern of response to beliefs and preferences differed across

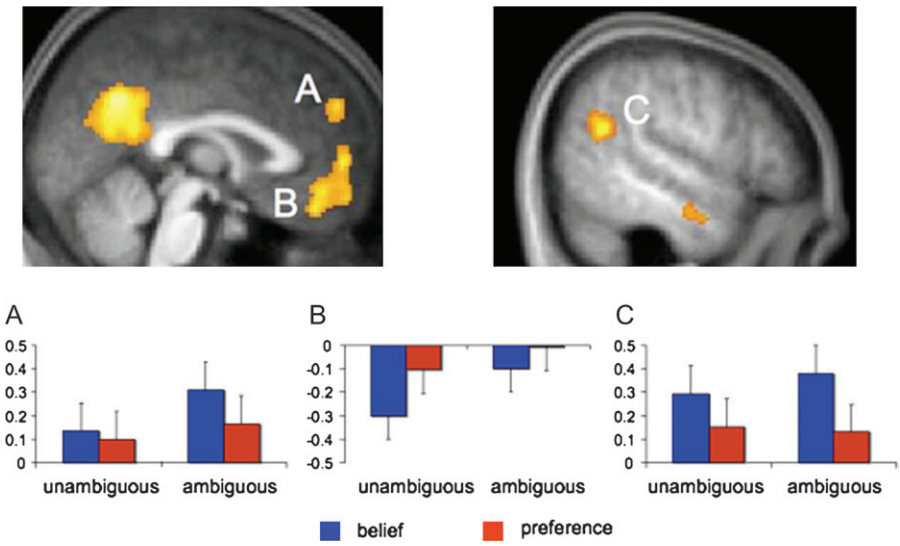

Figure 1. Average parameter estimates in dMPFC (*A*), vMPFC (*B*), and right TPJ (*C*) as a function of type of mental state (belief vs. preference) and mentalizing certainty (unambiguous vs. ambiguous). Activation in both dorsal and vMPFC was characterized by a main effect of certainty, such that both regions responded more during ambiguous than unambiguous inferences, regardless of content. In contrast, activation in right TPJ was characterized only by a main effect of content type, such that it responded more during inferences about beliefs than during inferences about preferences. Error bars represent confidence interval for within-subject designs (Loftus and Masson 1994).

these 3 regions; however, the 2-way interaction of region × ambiguity did not reach significance, $F_{2,42} = 1.79$, $P > 0.17$.

To confirm these findings, we also conducted a whole-brain, random-effects contrast of "ambiguous > unambiguous" scenarios. Consistent with the region-of-interest analysis, the sole region to emerge from this contrast was dMPFC. Additionally, whole-brain, random-effects contrasts of belief versus preference scenarios underscored the differential engagement of right TPJ and vMPFC as a function of content. Whereas right TPJ emerged from the contrast of "belief > preference," vMPFC emerged from the contrast of "preference > belief" (Table 3).

Discussion

The human ability to apprehend the mental states of others requires solutions to a host of cognitive challenges. The current findings add to the emerging empirical consensus that these challenges are met by an equally varied set of distinct cognitive processes rather than a single, monolithic "theory-of-mind" module. Replicating earlier research (Saxe and Kanwisher 2003; Saxe and Powell 2006), mentalizing about others' beliefs was associated with greater activity in right TPJ compared with mentalizing about others' preferences or to nonsocial processing. That understanding that others' beliefs would rely on such specialized processing has been anticipated by a number of commentators, who have pointed out that such inferences place unique demands on cognition, including a requirement to understand representational aspects of others' minds and to suspend attention to one's own knowledge in favor of understanding the unique knowledge possessed by another person (Apperly et al. 2005; Saxe 2006; Mitchell 2009).

In contrast, regardless of the type of mental state under consideration, both dorsal and ventral aspects of MPFC responded more during ambiguous, underspecified inferences than during unambiguous, well-constrained inferences. Comparisons across past studies have observed greater MPFC activation during relatively ambiguous inferences about preferences than during relatively unambiguous inference about beliefs, concluding that the relevant difference was in the type of mental state being considered (Van Overwalle 2009). However, the current results suggest a different conclusion. Here, dMPFC did not distinguish between inferences about beliefs and preferences when such inferences were matched for ambiguity, suggesting that what primarily drives the engagement of this region is not the type of mental state being inferred but rather the computational demands associated with constructing novel predictions from minimal information (Johnson-Laird 1994, 2001; Mitchell forthcoming).

What kinds of computational demands might these be? Recently, a number of commentators have suggested that MPFC contributes to a network of regions that subserves the construction of simulated scenarios. For example, in addition to its ubiquitous role in mentalizing, MPFC is consistently engaged by attempts to prospectively imagine the future and to retrospectively remember the past (Addis et al. 2007; Buckner and Carroll 2007; Schacter et al. 2007; Spreng et al. 2009), both of which require perceivers to use internally generated simulations of a situation that is divorced from the current context. Likewise, mentalizing under uncertainty may require perceivers to engage in similar processes of simulation, for example, by imagining their own response to an analogous situation or by drawing on aspects of comparable events from their own life. That is, when ambiguity about another person's mental states is high, our inferences about other minds may be guided by the contents of our own internal mental experience, mediated by MPFC (Mitchell, Banaji, Macrae 2005b; Mitchell et al. 2006; Jenkins et al. 2008).

Intriguingly, this observation suggests that one reason that MPFC has been so consistently associated with social cognition may be that inferences about the minds of other people are necessarily less constrained than inferences about the physical world. Because the mind of another person is inherently mutable and impossible to perceive directly, inferences about human minds may be fundamentally more ambiguous than inferences about our inanimate, physical surroundings (Mitchell forthcoming). To the extent that MPFC contributes to simulating plausible outcomes for indistinct and shifting phenomena, this region should be expected to participate frequently in understanding the minds of others.

However, such MPFC-mediated processes might also be engaged during nonsocial inferences that likewise require the consideration of multiple, "fuzzy" alternatives based on internally generated simulations. Humans must often make

Table 2

Peak voxel and number of voxels for brain regions obtained from the random-effects contrast of all mentalizing scenarios > nonsocial scenarios, $P < 0.05$, corrected

	x	*y*	*z*	Voxels
dMPFC	0	54	32	186
vMPFC	−8	50	−2	940
R superior temporal sulcus	56	−10	−20	133
L superior temporal sulcus	−68	−34	−4	215
R TPJ	54	−56	22	305
L TPJ	−48	−62	36	272
Medial parietal cortex	−8	−60	18	2477
R occipital cortex	12	−102	8	171
L occipital cortex	−24	−100	6	269

Note. Coordinates refer to the Montreal Neurological Institute stereotaxic space. R, right; L, left.

Table 3

Peak voxel and number of voxels for brain regions obtained from random-effects contrasts of certainty and type of mental state $P < 0.05$, corrected

	x	*y*	*z*	Voxels
Ambiguous > unambiguous				
MPFC	−4	36	40	241
Unambiguous > ambiguous				
No regions observed at $P < 0.05$, corrected				
Belief > preference				
R TPJ	50	−52	20	25
L TPJ	−50	−52	22	101
Preference > belief				
vMPFC	6	56	0	280
L orbitofrontal cortex	−22	36	−6	30
R insula	50	12	−8	162
	46	−10	−4	35
L inferior frontal gyrus	−46	2	14	30
Midcingulate cortex	−6	−8	30	32
Posterior cingulate cortex	0	−34	30	495
R intraparietal sulcus	30	−42	40	1739
L intraparietal sulcus	−30	−56	54	280
	−30	−36	40	82
	−46	−38	52	48
R middle temporal gyrus	58	−50	−10	111
Cerebellum	18	−64	−46	41
R lateral occipitotemporal sulcus	46	−60	−8	149
Superior parietal gyrus	−8	−78	40	52
Occipital cortex	4	−90	20	46

complex, underdetermined inferences outside the social domain, such as when deciding what kind of weather to expect during an upcoming trip or how the stock market will be affected by lower interest rates. In and of themselves, the current results cannot adjudicate whether MPFC contributions to uncertain inference making are limited to social situations (i.e., mentalizing) or may extend to relatively less social contexts. Indeed, recent findings demonstrate that regions of the right TPJ previously thought to be selective for social cognition also contribute to decidedly nonsocial tasks (Mitchell 2008; Scholz et al. 2009), raising the possibility that MPFC will also prove to participate across both social and nonsocial situations. The possibility that this region subserves processing of ambiguous information across multiple domains awaits future empirical test.

The Flexible Nature of Social Cognition

The current results also have implications for a longstanding debate over the question of how one person goes about "reading the mind" of another. Psychologists and philosophers have together posited 2 main accounts of the processes by which human beings understand other minds: broadly, those that are "simulationist" (Heal 1986; Gordon 1992) and those that are more "rule based" (also known as "theory" theories; Gopnik and Wellman 1994). Specifically, simulationist theories take as their starting point the observation that, although perceivers can never access the mind of another person directly, they do have constant and direct access to the conscious experience of one mind—their own—which they may be able to use as a model in which to understand the mental experience of another. Such theories suggest that, consciously or unconsciously, perceivers appeal to aspects of their own experience in order to generate insights into other minds. In contrast, rule-based theories emphasize the accumulation over one's lifetime of probabilistic laws about how human minds work (e.g., "people generally remember what they did 5 min ago"; "when people choose an object freely and consistently, they generally like that object"), which can be applied as relevant situations arise. Although simulationist and rule-based theories of social cognition have often been portrayed as mutually exclusive possibilities for how humans understand the minds of others, the current study suggests a more hybrid view. On one hand, rule-based mentalizing may be a useful strategy when perceivers reason about unambiguous mental states in ways that are strongly guided by explicit contextual information. However, more self-based simulationist processes may be needed to infer mental states under conditions of greater uncertainty, that is, when contextual cues less firmly constrain the possible goings on of another person's mind.

Interestingly, although both vMPFC and dMPFC differentiated between ambiguous and unambiguous inferences, vMPFC also showed greater activation during inferences about preferences than during inferences about beliefs. Analysis of participants' agreement on a single response for each vignette confirmed that preference scenarios were no more ambiguous than belief scenarios, and no other regions sensitive to ambiguity (e.g., dMPFC) distinguished between preferences and beliefs. As such, this finding replicates earlier studies that demonstrated greater response in vMPFC when mentalizing about others' affective states than their cognitive states (Hynes et al. 2006; Vollm et al. 2006; Shamay-Tsoory and Aharon-Peretz 2007) but suggests that this region may be sensitive not only to type of mental state being inferred but also the ambiguity of the information on which such an inference can be made (cf., Van Overwalle 2009).

A possible, albeit speculative, explanation for the less selective functional profile observed in vMPFC builds on social psychological research on attribution, which has long distinguished between explanations of behavior that focus on "the person" versus those that focus on "the situation" (Heider 1958). In the current study, unambiguous preference scenarios supported highly certain inferences because of what perceivers knew about their protagonists (i.e., the person), for example, that someone chose a particular item consistently despite having other options. In contrast, unambiguous belief scenarios supported highly certain inferences because they contained strong situational constraints, such that most human beings would be expected to believe the same thing under the same circumstances (Gilbert 1998), for example, that someone put an object in a particular place and returned to retrieve it a few minutes later. One possibility is that activity in vMPFC, which responded more during unambiguous inferences about preferences than during unambiguous inferences about beliefs ($P < 0.02$), could be associated specifically with person-focused attribution (Mitchell et al. 2005). This hypothesis raises the interesting possibility that vMPFC may respond more strongly to stable, idiosyncratic beliefs (such as a person's belief in ghosts or karma) that provoke high levels of person-based attribution than to transient preferences that depend heavily on the particular context (a person prefers mojitos to mimosas, but not before noon). Because the current study relied primarily on transient beliefs and stable preferences, additional research is needed to determine the specific contributions to mentalizing made by vMPFC, including its involvement in situation- versus person-based attribution.

Conclusion

The current findings continue the ongoing work of cleaving the functional neuroanatomy of social cognition into its constituent parts. Rather than comprising a single, monolithic process for contemplating the minds of others, recent research has increasingly made clear that social cognition decomposes into a number of distinct processes, each contributing some specific function to overall human social competence. Here, we suggest that one fruitful way to divide social cognition follows from the fact that perceivers face a number of uniquely different mentalizing challenges: not only the ability to infer a wide variety of mental states—such as beliefs, knowledge, feelings, and preferences—but also the ability to mentalize under varying degrees of uncertainty and ambiguity. The current results suggest that the human brain appears to respond to such demands by selectively engaging different regions as a function of the particular social–cognitive challenge to be met. Although some regions, such as the right TPJ, appear to contribute to social cognition by subserving inferences about specific types of mental states (i.e., beliefs), other regions, such as the dMPFC, are indifferent to the distinctions between others' beliefs and preferences. Instead, the MPFC may contribute preferentially to social cognition when making sense of new situations, unfamiliar individuals, or ambiguously motivated behavior—in other words, when mentalizing under uncertainty.

Funding

National Science Foundation (BCS 0642448); National Center for Research Resources (P41RR14075); Medical Investigation of Neurodevelopmental Disorders Institute.

Supplementary Material

Supplementary material can be found at http://www.cercor.oxfordjournals.org/

Notes

The authors thank Daniel L. Ames, Daniel Gilbert, Sean Loosli, and Lindsey Powell for advice and assistance. Data were collected at the Athinoula A. Martinos Center for Biomedical Imaging. *Conflict of Interest*: None declared.

Address correspondence to Adrianna C. Jenkins, Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02138, USA. Email: ajenkins@wjh.harvard.edu.

References

Addis DR, Wong AT, Schacter DL. 2007. Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. Neuropsychologia. 45:1363–1377.

Apperly IA, Samson D, Humphreys GW. 2005. Domain-specificity and theory of mind: evaluating neuropsychological evidence. Trends Cogn Sci. 9:572–577.

Buckner RL, Carroll DC. 2007. Self-projection and the brain. Trends Cogn Sci. 11:49–57.

Dennett DC. 1987. The intentional stance. Cambridge (MA): MIT Press.

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RS, Frith CD. 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition. 57:109–128.

Gallagher HL, Happe F, Frunswick N, Fletcher PC, Frith U, Frith CD. 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 HL, Jack AI, Roepstorff A, Frith CD. 2002. Imaging the intentional stance in a competitive game. Neuroimage. 16:814–821.

Gilbert DT. 1998. Ordinary personology. In: Gilbert DT, Fiske ST, Lindzey G, editors. Handbook of social psychology. New York: McGraw Hill. p. 89–150.

Goel V, Grafman J, Sadato N, Hallett M. 1995. Modeling other minds. Neuroreport. 6:1741–1746.

Gopnik A, Wellman HM. 1994. The theory theory. In: Hirschfeld LA, Gelman SA, editors. Mapping the mind: domain specificity in cognition and culture. New York: Cambridge University Press. p. 257–293.

Gordon RM. 1992. Folk psychology as simulation. Mind Lang. 1:158–171.

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

Heider F. 1958. The Psychology of interpersonal relations. New York: John Wiley & Sons.

Hynes CA, Baird AA, Grafton ST. 2006. Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. Neuropsychologia. 44:374–383.

Jenkins AC, Macrae CN, Mitchell JP. 2008. Repetition suppression of ventromedial prefrontal activity during judgments of self and others. Proc Natl Acad Sci USA. 105:4507–4512.

Johnson-Laird PN. 1994. Mental models and probabilistic thinking. Cognition. 50:189–209.

Johnson-Laird PN. 2001. Mental models and deduction. Trends Cogn Sci. 5:434–442.

Kelley HH. 1972. Attribution in social interaction. In: Jones EE, Kanouse DE, Kelley HH, Nisbett RE, Valins S, Weiner B, editors. Attribution: perceiving the cause of behavior. Hillsdale (NJ): Lawrence Erlbaum and Associates. p. 1–26.

Loftus GR, Masson MEJ. 1994. Using confidence intervals in within-subject designs. Psychon Bull Rev. 1:476–490.

Mitchell JP. 2008. Activity in right temporo-parietal junction is not selective for theory-of-mind. Cereb Cortex. 18:262–271.

Mitchell JP. 2009. Inferences about other minds. Philos Trans R Soc Lond B Biol Sci. 364:1309–1316.

Mitchell JP. Forthcoming. Social psychology as a natural kind. Trends Cogn Sci.

Mitchell JP, Banaji MR, Macrae CN. 2005a. General and specific contributions of the medial prefrontal cortex to knowledge about mental states. Neuroimage. 28:757–762.

Mitchell JP, Banaji MR, Macrae CN. 2005b. The link between social cognition and self-referential thought in the medial prefrontal cortex. J Cogn Neurosci. 17:1306–1315.

Mitchell JP, Heatherton TF, Macrae CN. 2002. Distinct neural systems subserve person and object knowledge. Proc Natl Acad Sci USA. 99:15238–15243.

Mitchell JP, Macrae CN, Banaji MR. 2005. Forming impressions of people versus inanimate objects: social-cognitive processing in the medial prefrontal cortex. Neuroimage. 26:251–257.

Mitchell JP, Macrae CN, Banaji MR. 2006. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron. 50:655–663.

Samson D, Apperly IA, Chiavarino C, Humphreys GW. 2004. Left temporoparietal junction is necessary for representing someone else's belief. Nat Neurosci. 7:499–500.

Saxe R. 2006. Uniquely human social cognition. Curr Opin Neurobiol. 16:235–239.

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

Saxe R, Powell LJ. 2006. It's the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci. 17:692–699.

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

Schacter DL, Addis DR, Buckner RL. 2007. Remembering the past to imagine the future: the prospective brain. Nat Rev Neurosci. 8:657–661.

Scholz J, Triantafyllou C, Whitfield-Gabrieli S, Brown EN, Saxe R. 2009. Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. PLoS ONE. 4, doi:10.1371/journal.pone.0004869.

Shamay-Tsoory SG, Aharon-Peretz J. 2007. Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. Neuropsychologia. 45:3054–3067.

Spreng RN, Mar RA, Kim AS. 2009. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J Cogn Neurosci. 21:489–510.

Van Overwalle F. 2009. Social cognition and the brain: a meta-analysis. Hum Brain Mapp. 30:829–858.

Vollm BA, Taylor AN, Richardson P, Corcoran R, Stirling J, McKie S, Deakin JF, Elliott R. 2006. Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a non-verbal task. Neuroimage. 29:90–98.

Zaitchik D. 1990. When representations conflict with reality: the preschooler's problem with false beliefs and "false photographs". Cognition. 35:41–68.

OPEN ACCESS Freely available online

PLOS ONE

The Neural Bases of Directed and Spontaneous Mental State Attributions to Group Agents

Adrianna C. Jenkins[1]*, David Dodell-Feder[2], Rebecca Saxe[3], Joshua Knobe[4]

1 Helen Wills Neuroscience Institute and Haas School of Business, University of California, Berkeley, California, United States of America, **2** Department of Psychology, Harvard University, Cambridge, Massachusetts, United States of America, **3** Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **4** Program in Cognitive Science, Yale University, New Haven, Connecticut, United States of America

Abstract

In daily life, perceivers often need to predict and interpret the behavior of group agents, such as corporations and governments. Although research has investigated how perceivers reason about individual members of particular groups, less is known about how perceivers reason about group agents themselves. The present studies investigate how perceivers understand group agents by investigating the extent to which understanding the 'mind' of the group as a whole shares important properties and processes with understanding the minds of individuals. Experiment 1 demonstrates that perceivers are sometimes willing to attribute a mental state to a group as a whole even when they are not willing to attribute that mental state to any of the individual members of the group, suggesting that perceivers can reason about the beliefs and desires of group agents over and above those of their individual members. Experiment 2 demonstrates that the degree of activation in brain regions associated with attributing mental states to individuals—i.e., brain regions associated with mentalizing or theory-of-mind, including the medial prefrontal cortex (MPFC), temporo-parietal junction (TPJ), and precuneus—does not distinguish individual from group targets, either when reading statements about those targets' mental states (directed) or when attributing mental states implicitly in order to predict their behavior (spontaneous). Together, these results help to illuminate the processes that support understanding group agents themselves.

Citation: Jenkins AC, Dodell-Feder D, Saxe R, Knobe J (2014) The Neural Bases of Directed and Spontaneous Mental State Attributions to Group Agents. PLoS ONE 9(8): e105341. doi:10.1371/journal.pone.0105341

Editor: Allan Siegel, University of Medicine & Dentistry of NJ - New Jersey Medical School, United States of America

Received May 5, 2014; **Accepted** July 17, 2014; **Published** August 20, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files. The full fMRI dataset is available upon request.

Funding: This work was supported by a John Merck Scholars Program award to RS (http://www.jmfund.org/jm_scholars_program.php). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: adrianna.jenkins@berkeley.edu

Introduction

In domains ranging from the economy to national security, large-scale decisions often involve judgments about the machinations of a group agent, such as a terrorist organization, government, or corporation. Sometimes, judgments about a group agent simply reduce to judgments about one or more of its individual members (for example, thinking about whether or not a *country* is hiding nuclear weapons may primarily involve consideration of that country's *leader*). However, people also sometimes appear to make judgments about a group by treating it as an entity in and of itself. Individuals assign moral blame and punishment to whole organizations [1], interpret laws by looking for the 'intentions' of the legislature [2], may get into financial trouble by reasoning about the 'mind' of the market [3], and, in a recent decision by the United States Supreme Court, extended rights typically granted to individuals to a corporation as a whole [4].

Although an abundance of research has investigated the effects of group membership on how people perceive and reason about the minds of individual people (for recent reviews, see [5–7], less is known about how perceivers reason about the 'mind' of a group agent itself [8]. To investigate this question, the present work uses a combination of behavioral and fMRI approaches to examine the extent to which understanding the 'mind' of the group as a whole shares important properties and processes with understanding the minds of individuals. Specifically, we ask (1) to what extent people sometimes reason about the beliefs and intentions of a group agent separately from those of the groups' members and (2) to what extent brain regions associated with understanding individuals also support understanding group agents.

In order to predict or understand the behavior of a single individual, perceivers often appeal to that individual's *mental states* (i.e., his or her thoughts, beliefs, intentions, desires, and feelings). This capacity to ascribe mental states to others—that is, to *mentalize* [9,10] or engage *theory-of-mind* [11,12]—reveals itself in the words perceivers use when talking about other people. For example, we can say that Dick *thought* he was aiming for a partridge and never *intended* to shoot his friend. Words like *think, believe, feel, intend, want*, and *plan* all refer to the inner contents of other minds, allowing perceivers to speak about the purported underlying causes of others' behavior even as they diverge from that behavior itself [13,14]. In turn, inferences about these internal causes guide moral decisions about how others should be treated, including the extent to which they deserve praise or punishment [15,16].

Over the past two decades, an abundance of neuroimaging research has linked mentalizing or theory-of-mind to a consistent

set of brain regions, including the medial prefrontal cortex (MPFC), temporo-parietal junction (TPJ), and precuneus/posterior cingulate, sometimes collectively called the 'theory-of-mind network'. Using carefully controlled tasks that aim to isolate theory-of-mind, these regions show preferential engagement when people are thinking about humans versus other entities [17–24] and when people are thinking about humans' minds versus their other aspects, such as their physical attributes [21,25–27]. Although much of this evidence has been correlational, recent work using TMS has demonstrated a causal role for the Right TPJ (RTPJ) in the use of mental state information for moral judgment [15], and research on individuals with damage to MPFC and TPJ has demonstrated a role for those regions in the ability to make inferences about others' mental states [28,29].

Intriguingly, mental state words pervade perceivers' statements not only about individuals but also about groups. In recent news reports, we learn that "Apple thinks carefully about its entire product lineup" [30], that "Apple wants owners to sell their old iPhones back to the company for a discount on a new phone" [31], and that "Apple intends to work with record labels to identify and promote up and coming artists" [32]. In cases like these, people apply words normally associated with the psychological states of an individual person—words like 'thinks', 'wants', and 'intends'—to a corporation as a whole. These same expressions can also be applied to other sorts of group agents. People talk about what a government agency 'intends', what a religious organization 'thinks', or what a sports team 'loves' or 'hates' [33–37]. Indeed, archival studies show that people speak about groups using mental state words spontaneously, even outside the context of an experiment [36], and cross-cultural studies document the use of mental state words in descriptions of groups not only in the West, but also in East Asian cultures [35,37].

Does the use of such language indicate that people understand governments and other organizations by attributing mental states to a group? Critically, there are two different senses in which one might think about 'groups' and, accordingly, two different senses in which one might investigate the processes perceivers use to understand groups. On one hand, one could think about a 'group' as referring to the *members* of groups. If each group member is a human being, then the group is simply a collection of human beings. A first sense in which one might investigate how perceivers understand groups, then, is to investigate how people understand collections of human beings. On the other hand, one could think about a 'group' as referring to a *group agent* [38,39]. A group agent itself is not merely a collection of separate human beings but, instead, an entity with whatever sort of status attaches itself to corporations, nations, and sports teams. Thus, a second sense in which one might investigate how perceivers understand groups is to investigate how people understand not collections of individuals, but group agents.

An example highlights the distinction between a group in the sense of a collection of individuals and a group in the sense of a group agent. Consider the sentence "The employees and stockholders of Acme Corp. are all in debt." This sentence says something about the financial condition of various individual human beings while making no claims about the financial condition of the corporation with which they are associated. In other words, the sentence ascribes a property to the members without ascribing that property to the group agent itself. By contrast, consider the sentence, "Acme Corp. is in debt." This sentence says something about the financial condition of a corporation, but it makes no claims at all about the financial condition of any individual human beings. (The corporation itself could be in debt even if all of the employees and stockholders were in excellent financial shape.) Thus, this sentence ascribes a property to a group agent without ascribing that same property to any of the members.

Existing work already provides some evidence for the claim thinking about groups in the first sense—i.e., thinking about collections of human beings—shares properties and processes with thinking about individual people. Behaviorally, the vast literatures on stereotypes and intergroup relations show that people are willing to ascribe psychological attributes to whole collections of others [7,40–45], and studies indicate that some of the same principles that apply to the ascription of properties to individual agents also appear in the ascription of properties to whole collections of agents [46,47]. Moreover, a recent neuroimaging study observed activation in brain regions associated with theory-of-mind—MPFC, TPJ, and precuneus—when participants evaluated the applicability of certain preferences both to individual people and to collections of individuals, compared to a non-mental control condition [48]. Taken together, these behavioral and neuroimaging studies provide support for the view that people can ascribe psychological attributes not only to individual human beings but also to collections of human beings, and that they may use similar processes to do so (even if the outcomes of those processes may sometimes differ [47,49]).

Yet studies like these still leave open the question of how people understand groups in the second sense—i.e., how they understand group agents. As we saw above, people can ascribe a non-mental property to all of the members of a group agent without ascribing that property to the group agent itself ("All of the employees and stockholders are in debt"). Similarly, perhaps people can ascribe a mental property (i.e., a mental state) to all of the members of a group without in any way ascribing these states to the group agent itself ("The employees and stockholders all love Jeopardy!"). We have also seen that people can ascribe a non-mental property to a group without ascribing that property to the individual members ("Acme Corp. is in debt."). Similarly, perhaps people can ascribe mental states to a group agent without ascribing that state to any of the members. Indeed, recent research suggests that the more people perceive a 'group mind', the less they tend to perceive the minds of the members of that group [8,50].

With this in mind, the current studies investigate how perceivers understand group agents by examining the extent to which understanding group agents shares important properties and processes with understanding individuals. Experiment 1 examines behaviorally the extent to which people ascribe mental states to group agents over and above attributions of mental states to their individual members. Experiment 2 uses fMRI to investigate the extent to which understanding and predicting the behavior of group agents recruits brain regions associated with understanding and predicting the behavior of individuals—i.e., brain regions associated with theory of mind.

Experiment 1: Ascriptions to group agents vs. ascriptions to group members

When people use sentences that appear to ascribe mental states to a group agent, are they actually ascribing something to the group agent, or are they merely attributing something to the group's members? For example, consider the sentence, "United Food Corp. believes that the new policy is morally unacceptable." At least on the surface, this sentence appears to attribute a mental state (the belief that the policy is morally unacceptable) to a group agent (United Food Corp). However, it is possible that this is just a linguistic shortcut, and that when people use or hear sentences like

this one, they are really attributing mental states to the members of the group, not to the group itself.

Existing research demonstrates that people sometimes do use sentences that appear to attribute a property to a group when referring to its members, specifically when the members of the group have the particular property in their roles as group members [39]. For example, if each member of the Sigma Chi fraternity gets drunk, and if each of them does so in his role as a Sigma Chi member, people tend to agree with the sentence, "The Sigma Chi fraternity got drunk" [39]. This sentence appears on the surface to be ascribing a property to the fraternity itself—the actual organization— but is in fact just a shorthand way of ascribing a property to the individual members in their roles as members.

In Experiment 1, we examine whether apparent mental state attributions to group agents can involve attributions of a property to a group agent itself, or whether they reduce to attributions to individual group members. To the extent that perceivers genuinely attribute a property to the group agent itself, attributions to group agents should sometimes diverge from attributions to the members of those groups. That is, we should observe (a) cases in which perceivers attribute a mental state to all of the members of the group without attributing that state to the group agent itself and (b) cases in which perceivers attribute a mental state to the group agent without attributing that state to any of the group's members. In contrast, to the extent that apparent attributions to group agents are merely shorthand for attributions to the group members, participants should not attribute properties to the group agent that they do not also attribute to the members of the group. Thus, finding that individuals attribute mental states to a group agent without attributing that state to any of the group's members would be the most unambiguous evidence that perceivers can apply mental states to group agents themselves.

Method

Participants. 116 Yale students and faculty (33% female; age range 18-54, mean age 21 years) were recruited outside a dining hall to fill out a questionnaire for payment.

Ethics statement. This study was approved by the Institutional Review Board at Yale University. All participants provided written informed consent.

Materials and Procedure. This experiment used a 2 (mental state: individual-only or group-only) × 3 (question: any member, each member, group) design in which target was manipulated within-subject and question type was manipulated between subjects. Each participant received eight vignettes in counterbalanced order. Four vignettes were designed in such a way that it would be logically possible to ascribe a particular mental state to each of the individuals in the group without ascribing that state to the group itself (*Individual-only* condition). For example, one vignette described an organization devoted to fighting the death penalty. All of the members of this anti-death penalty organization are also interested in antebellum American history, so they decide to form a separate organization, with exactly the same members, called the Shady Grove Antebellum Historical Society (SGAHS), which meets to discuss historical questions. If participants are willing to ascribe a mental state to all of the individual members without ascribing that mental state to the group as a whole, participants should report that all of the members of SGAHS want to fight the death penalty but that the SGAHS itself *does not* want to fight the death penalty. On the other hand, to the extent that attributions to a group simply reduce to the attributions made to the individual members, participants should report that SGAHS *does* want to fight the death penalty.

The other four vignettes were designed such that that it would be logically possible to ascribe a mental state to the group itself without ascribing that state to any of the individual members (*Group-only* condition). For example, one vignette described a large organization that was commissioned to build a space shuttle. Some members of the organization put together the software, others build the exterior, still others are in charge of the fuel, and so forth. But there is no single person who works on every aspect of the project. To the extent that people are willing to ascribe a property to a group agent over and above its members, participants should say that the organization knows how to build a space shuttle, but the individual members do not. In another vignette, a Community Association needs to choose music for an upcoming event. Some members really want to play punk music and can't stand classical, others really want to play classical music but strongly dislike punk, so in the end, the Association selects a third option: classic rock. If people are willing to attribute properties to group agents over and above their members, participants should say that the Community Association itself preferred playing classic rock but that none of the individual members shared this preference. On the other hand, to the extent that attributions to the group simply reduce to the attributions made to the individual members, participants should report either that most or all of the individual members prefer playing classic rock or that the group itself does not prefer playing classic rock. For full texts of the vignettes, see (Text S1).

Each participant was randomly assigned to one of three question conditions: 'any member,' 'each member,' or 'group.' Participants in the 'any member' condition received after each vignette a question about whether any individual member of the group had a particular mental state ('Do any of the members of the Community Association prefer the idea of playing classic rock to the idea of playing every other type of music?'). Participants in the 'each member' condition were asked whether *each* member had the relevant state ('Do each of the individual members of the Community Association prefer…?'). Finally, participants in the 'group' condition received questions about whether the group itself had the relevant state ('Does the Community Association prefer…?'). Each question was answered on a scale from 1 ('No') to 7 ('Yes').

Results

Two participants failed to complete all items of the questionnaire. We calculated the mean response to 'group', 'any member', and 'each member' questions in the 'Members Only' vignettes and the 'Group Only' vignettes for the remaining participants (see Fig. 1). To the extent that participants attributed purported mental states to group agents themselves, we should observe both cases in which participants attribute a state to all of the members of the group without attributing that state to the group itself and, most critically, cases in which participants attribute a state to the group itself without attributing that state to any of the individual members. See (Table S1) for complete dataset.

For the Members-Only vignettes, a one-way ANOVA revealed a significant effect of question condition, $F(2, 114) = 41.2$, $p < .001$, $\eta^2 = .42$ (Fig. 1), such that participants were willing to attribute states to some or all of the members of a group without attributing those states to the group itself. Tukey's posthoc tests showed that participants agreed less with ascriptions in the 'group' question condition than in either the 'any member' question condition, $p < .001$, or the 'each member' question condition, $p < .001$, suggesting that attributions to the group did not simply reduce to attributions to the group's members.

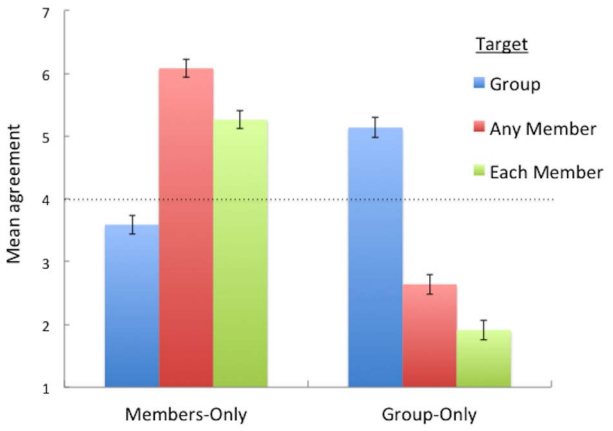

Figure 1. Mean agreement with mental state ascriptions by condition for the Members-Only and Group-Only vignettes. Error bars show SE mean. Dotted black line indicates neutral midpoint; points above indicate agreement and points below indicate disagreement.
doi:10.1371/journal.pone.0105341.g001

Critically, for the Group-Only vignettes, a one-way ANOVA again revealed a significant effect of question condition on participants' responses, $F(2, 114) = 91.6$, $p < .001$, $\eta^2 = .62$ (Fig. 1), such that participants were willing to attribute states to the group itself that they did not attribute to any of the members of the group. Tukey's posthoc tests showed that participants agreed more with ascriptions in the 'group' question condition than in either the 'any member' question condition, $p < .001$, or the 'each member' question condition, $p < .001$. Moreover, participants' responses in the group question condition were significantly above the neutral midpoint of the scale, $p < .001$, indicating that participants were genuinely endorsing sentences ascribing mental states to group agents. These results suggest that attributions to the group agent were made over and above the attributions made to individual members.

This study explored the relationship between ascribing states to group agents and their members. We observed cases in which participants attributed a state to all of the members but did not attribute that state to the group itself and also cases in which participants attributed a state to the group itself but did not attribute the state to any of the members. Together, these results demonstrate that mental state ascriptions to a group agent can diverge from those made to the group's individual members, suggesting that perceivers can attribute a property of some sort to the group agent itself.

Experiment 2: Neural processes supporting mental state ascriptions to group agents

Experiment 1 suggests that that when people use expressions of the form 'United Food Corp. wants.', they appear to be ascribing something to the group itself, rather than to the members of the group. However, a further question concerns the processes supporting these ascriptions. That is, although such statements clearly involve the same linguistic expressions that people use when applying theory-of-mind to individual human beings, to what extent do they also involve the same cognitive processes?

To investigate the processes supporting attributions of purported mental states to group agents, we scanned participants using fMRI as they considered the mental states of individuals and groups. In one task, participants read sentences that referred explicitly to the mental states of groups and individuals (along with matched, non-mental control sentences). In a second task, participants carried out a procedure that relied on mental state ascription incidentally, without the use of mental state words: making predictions about what an individual or group would do in a variety of situations. To the extent that perceivers rely on processes associated with understanding individuals when they understand and predict the behavior of groups, brain regions associated with theory-of-mind should be active both when thinking about individuals and when thinking about group agents, and they should be active to a similar degree. On the other hand, to the extent that perceivers rely on different processes to understand group agents, we should observe reduced activation in brain regions associated with theory-of-mind—RTPJ, MPFC, and precuneus—during consideration of groups versus individuals. In the design of this study, steps were taken to (a) minimize, as much as possible, the likelihood that participants would simply consider the minds of individual group members when considering group agents and (b) test sensitively the degree to which brain regions associated with theory of mind are engaged during consideration of group agents. Unlike past studies, no individuals were mentioned or shown in the group condition, and both directed and spontaneous theory of mind tasks were included. Moreover, the results of Experiment 1 show that perceivers do interpret sentences about group mental states as ascribing mental states to the group agent itself.

Although MPFC, TPJ, and precuneus have all been associated consistently with theory-of-mind, finer-grained differences in the response profiles of these regions facilitate predictions about their involvement during consideration of group agents. Recent neuroimaging research has increasingly revealed that, even when mental state attributions to individuals are concerned, MPFC, TPJ, and precuneus do not all respond in the same ways under the same circumstances. In particular, there are at least two ways in which the processes associated with purported mental state reasoning about group agents may differ from those associated with individual people. One is that certain properties of the *type* of mental state content being attributed may differ. The other is that certain properties of the *target* to whom that content is being attributed may differ.

The RTPJ consistently demonstrates sensitivity to the *type* of mental state being ascribed. Specifically, a series of studies has demonstrated that RTPJ is selective for processing representational mental states, such as beliefs [51–55]; see [56] for review. The RTPJ response is high when participants read stories that describe a character's true or false beliefs but low during stories containing other socially salient information, such as a character's physical appearance, cultural background, or even internal sensations such as hunger or fatigue [25]. Similarly, activation in RTPJ is higher during inferences about an individual's beliefs than during closely matched inferences about an individual's preferences regardless of whether such inferences are more or less constrained by external information—a response profile that is not shared by other regions associated with social cognition, such as MPFC [57]. Moreover, activation in the RTPJ consistently tracks with thinking about mental contents, not merely seeing mental state words. RTPJ becomes engaged when participants think about others' mental states even in the absence of any mental state words, such as when they view non-verbal cartoons [58] or read descriptions of actions that imply a particular mental state [22]. Conversely, mental state words alone do not elicit activation in the RTPJ; for review see [59]. Thus, mental state words are neither necessary nor sufficient for eliciting RTPJ activation. Instead,

RTPJ activation during social cognition appears to be associated with the ascription of representational mental state content; for discussion see [60–62]. Thus, to the extent that perceivers attribute representational mental states to group agents, we should observe similar levels of RTPJ activation during consideration of group agents and individuals, both of which should exceed that associated with a non-mental control condition.

In contrast, MPFC appears to be especially sensitive to the *target* of mental state ascription. In particular, thinking about oneself, a similar individual, a familiar individual, or an individual whose perspective one has taken earlier is associated with more MPFC activation than thinking about more distant others [63–67]. MPFC also appears to be sensitive to the target of consideration when theory-of-mind is not explicitly called for. For example, this region exhibits less activation during consideration of "dehumanized" than "humanized" individuals [68] and responds more during consideration of one's own versus another person's physical attributes [26]. Although it remains open to further inquiry whether lower MPFC response in these cases genuinely indexes a difference in the degree to which mental states are attributed [68] or rather the use of an alternative process for doing so [57,63,67], the sensitivity of MPFC to the target of judgment suggests that group agents may be particularly likely to be associated with lower activation than individuals in this region.

Method

Participants. Nineteen right-handed, native English speakers (10 female; age range 19-25, mean age 21 years) with no history of neurological problems participated for payment. All participants had normal or corrected-to-normal vision.

Ethics statement. This study was approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) at the Massachusetts Institute of Technology. All participants provided written informed consent.

Stimuli and Behavioral Procedure. ***Directed theory-of-mind task.*** During fMRI scanning, participants completed an individual vs. group agent theory-of-mind task in which they read short statements about everyday events. Participants were instructed to read each statement and were told that they would be asked a series of questions about the statements later on in the experiment. Inanimate (control) statements communicated information without reference to people (e.g., "Although there wasn't much real data on agricultural production, the statistics showed that rutabaga production was consistently going down."). Based on each control statement, an *individual* statement and a *group* statement were constructed. *Individual* statements concerned a single person's mental state (e.g., "Although there wasn't much real data on agricultural production, George Hailwood was sure that rutabaga production was going down."). *Group* statements concerned the 'mental state' of a group agent (e.g., "Although there wasn't much real data on agricultural production, United Food Corp. was sure that rutabaga production was going down."). No participant viewed more than one version of the same base statement.

In each run of this task, participants read statements organized around a single theme (e.g., one run concerned George Hailwood, United Food Corp., and food production, whereas another concerned Stephanie Ann Majors, a record company, and music sales). For full texts of the stimuli, see (Text S2). Participants completed ten functional runs of eighteen statements each (six per condition), totaling 180 trials. Statements were displayed in random order within each run and remained onscreen for 8 s. Trials were separated by a variable inter-stimulus interval (2–16 s) during which participants passively viewed a black screen.

Spontaneous theory-of-mind task. Following each run of the directed theory-of-mind task, participants were asked to make a series of predictions about the individual and group about which they had just read (e.g., "The asparagus might be contaminated by bacteria. Would George Hailwood [United Food Corp.] be more likely to (a) recall all of the asparagus or (b) cover up the whole incident?"). This task elicited mental state reasoning indirectly by asking participants to formulate predictions about behavior, such that no mental state words were presented to participants at any point. Each question remained onscreen for 12 s, and participants were obliged to respond during that time by pressing one of two buttons on a button box held in the left hand. Each run comprised eight trials (four per condition) separated by 10 s. Each participant answered each question either for the individual or the group, but not both (question assignment randomized across participants).

Theory-of-mind localizer. In order to facilitate region-of-interest (ROI) analyses focusing on brain regions associated with theory-of-mind, participants also completed a functional localizer task in which they read short narratives and made inferences about individual protagonists' beliefs (e.g., concerning the location of a hidden object) and inferences about physical representations (e.g., the contents of an outdated photograph [22]). Each narrative was displayed for 10 s and was followed by a statement that participants judged as true or false (e.g., Belief story: "Sarah thinks her shoes are under the dress"; Physical story: "The original photograph shows the apple on the ground") which remained onscreen for 4 s. Participants were obliged to respond during that time by pressing one of two buttons. Trials were separated by 12 s fixation. Participants completed four runs, each of which comprised eight trials (four per condition), for a total of 32 trials.

Imaging Procedure. fMRI data were collected using a 3 Tesla Siemens scanner. Functional imaging used a gradient-echo echo-planar pulse sequence (TR = 2 s; TE = 30 ms; flip angle = 90°, 30 near-axial slices, 4 mm thick, in-plane resolution = 3×3 mm, whole brain coverage). These sequences used PACE online motion correction for movement < 8 mm. fMRI data were preprocessed and analyzed using SPM2 (Wellcome Department of Cognitive Neurology, London, United Kingdom) and custom software. Data from each subject were motion corrected and normalized into a standard anatomical space based on the ICBM 152 brain template (Montreal Neurological Institute). Normalized data were then spatially smoothed (5 mm full-width-at-half-maximum [FWHM]) using a Gaussian kernel.

Statistical analyses were performed using the general linear model in which the event-related design was modeled using a canonical hemodynamic response function and other covariates of no interest (a session mean and a linear trend). After these analyses were performed individually for each participant, the resulting contrast images for each participant (i.e., *individual* > *control, group* > *control*) were entered into a second-level analysis in which participants were treated as a random effect. Data were thresholded at $p<.001$, $k>10$, uncorrected.

For the directed theory of mind task, conjunction analysis was performed following the procedure described by Cabeza, Dolcos, Graham, & Nyberg [69]. Whole-brain statistical maps were created from the *individual* > *control* and *group* > *control* contrasts separately to identify voxels activated by each condition (thresholded individually at $p < .01$), making for a conjoint threshold of $p < .001$.

ROIs were defined for each subject individually based on a whole-brain analysis of the theory-of-mind localizer in three regions: RTPJ, precuneus, and MPFC. Regions were defined as 10 or more contiguous voxels that were significantly more active ($p < 0.001$, uncorrected) during stories about mental states than during

control stories about physical representations. The average responses relative to rest during the individual and group conditions were then estimated in these regions. Within each ROI, the mean percent signal change (PSC = 100 × raw BOLD magnitude for [condition − rest]/raw BOLD magnitude for rest) was calculated for each condition at each time point (averaging across all voxels in the ROI and all trials of the same condition) and averaged across seconds 6–10 to account for hemodynamic lag. Individual subject means for each condition of each task are available as (Table S2). The full fMRI dataset is available upon request.

Results

Directed theory-of-mind task. In order to assess the extent to which common cognitive processes subserve thinking about the minds of individuals and groups, we first conducted whole-brain, random effects analyses of BOLD signal. In whole-brain analyses, activation when participants contemplated the mental states of both individuals and groups (compared to control) was observed in brain regions associated with theory-of-mind, including MPFC, RTPJ, and precuneus. The direct comparisons between the individual and group conditions (individual <> group) yielded no areas of differential activation in regions typically associated with social cognition (Table 1). To the extent that overlapping BOLD activation reflects the engagement of overlapping cognitive processes, these initial observations suggest that thinking about individuals and groups may draw upon shared theory-of-mind processes.

Next, to test more directly the extent to which overlapping regions of cortex were recruited during contemplation of the mental states of individuals and groups, we conducted a conjunction analysis on the individual > control and group > control contrasts. This analysis revealed conjoint activation specifically in brain regions associated with theory-of-mind–MPFC, right and left TPJ, and precuneus–suggesting further that thinking about individuals and groups draw upon shared processes (Table 2; Fig. 2).

Although the foregoing analyses suggest that similar processes subserve thinking about individuals and groups as compared to a control condition, they leave open the possibility that thinking about individual and group agents may recruit theory-of-mind processes to different degrees. In order to evaluate the degree to which processes associated with theory-of-mind were recruited when thinking about individuals versus groups, we conducted independent region-of-interest (ROI) analyses within the regions of MPFC, RTPJ, and precuneus identified by the independent theory-of-mind localizer. Because the mental states in the localizer task were attributed to individual protagonists, this analysis technique provides a particularly stringent test for whether thinking about group agents genuinely recruits processes associated with thinking about individuals. Consistent with previous research, the theory-of-mind localizer (belief > photo contrast) yielded activation in MPFC (17/19 participants), RTPJ (19/19 participants), and precuneus (19/19 participants); Fig. 3. First, ROI analyses of the main task confirmed that each of these regions showed greater activation in the individual condition than in the control condition (MPFC, $t(16) = 2.28$, $p < .04$, $d = 0.57$; Right TPJ, $t(18) = 2.43$, $p < .03$, $d = 0.57$; precuneus, $t(18) = 5.99$, $p < .0001$, $d = 1.41$). Second, ROI analyses further revealed that each of these regions showed greater activation in the group condition as compared to control (MPFC, $t(16) = 2.22$, $p < .04$, $d = 0.55$; Right TPJ, $t(18) = 2.39$, $p < .03$, $d = 0.56$; precuneus, $t(18) = 6.32$, $p < .0001$, $d = 1.49$). Finally, no significant differences were observed between the responses to individuals versus groups in any of these regions, (MPFC, $t(16) = 0.69$, $p = .5$; Right TPJ, $t(18) = 0.09$, $p = .93$; precuneus, $t(18) = 1.51$, $p = .15$; Fig. 3). Together, these analyses suggest that brain regions associated with theory-of-mind are recruited to a highly similar degree during the contemplation of individuals and groups.

Spontaneous theory-of-mind task. The design of the previous task raises the possibility that activation during the individual and group conditions may have differed from the control condition due to the explicit use of mental state words (e.g., thinks, believes, wants) in the individual and group conditions. To explore whether common theory-of-mind processes subserve attributions to individuals and groups even when no mental state terms are used, we analyzed data from the portion of the study during which participants made predictions about the behavior of individuals and groups. Specifically, we compared activation during the individual and group conditions of the prediction task in the same regions of RTPJ, MPFC, and precuneus identified by the theory-of-mind localizer. Results replicated those from the directed theory-of-mind task. Consistent with the hypothesis that thinking about the minds of individuals and groups recruit similar theory-of-mind processes, activations above baseline were observed across the network in both the individual, $t(19) = 2.84$, $p < .02$, $d = 0.65$, and the group condition, $t(19) = 2.23$, $p < .04$, $d = 0.51$ (averaging across regions), and no differences were observed between the individual and group conditions in RTPJ ($M_{ind} = -.004$ $M_{group} = -.019$, $t(19) = 0.86$, $p > .39$), MPFC ($M_{ind} = .197$ $M_{group} = .180$, $t(19) = 0.36$, $p > .72$), or precuneus ($M_{ind} = .266$ $M_{group} = .231$, $t(19) = 1.64$, $p > .12$). For individual subject data, see (Table S2). These results suggest that the similar patterns of activation in the individual and group conditions observed in the first task are not simply due to the common use of mental state terms in those conditions. Here, when no mental state terms were presented, making predictions about individual and group agents' behavior also recruited the theory-of-mind network to an indistinguishable degree.

Discussion

In describing corporations, government agencies and other organizations, people sometimes use sentences of the form 'Apple thinks…' or 'The CIA wants…' The aim of the present investigation was to help illuminate how people think about group agents. The results of Experiment 1 indicate that sentences like these are ascribing something to the group agent itself. Perceivers used expressions like 'believes' and 'wants,' not merely to talk about some or all of the individual members of a group, but to talk about the group agent. Thus, attributions to the group sometimes diverged from attributions to the individual members: participants were willing to attribute a state to the group itself even when they were not willing to attribute that state to any of the individual members, and they were willing to attribute a mental state to all members of a group even when they were not willing to attribute that state to the group itself. In turn, the results of Experiment 2 reveal that that such ascriptions recruit brain regions associated with thinking about the minds of individuals, i.e., brain regions associated with theory-of-mind, both when theory-of-mind use is called for explicitly and when it arises spontaneously.

Past research has demonstrated consistent engagement of a particular network of regions, including MPFC, RTPJ, and precuneus, during inferences about the minds of individual people, i.e., during theory-of-mind. Across two tasks, we observed activation in this network when participants read or made predictions about group agents. In the *directed theory-of-mind task*, participants read about the states of individuals, group agents, and inanimate objects. In the *spontaneous theory-of-mind* task,

Table 1. Regions emerging from whole brain analyses.

Region	*x*	*y*	*z*	*T* value
Theory-of-mind Localizer (Belief > Photo)				
PC	2	−64	42	10.73
Right TPJ	58	−54	34	6.38
MPFC	0	52	46	6.27
Right STS	56	−26	−10	5.74
Left TPJ	−48	−52	20	5.39
Left Anterior STS	−54	4	−24	5.00
Left STS	−54	−20	−14	4.64
Right Temporal Pole	54	6	−34	4.51
Left Temporal Pole	−36	16	−26	4.39
Individual > Control				
PC	−6	−68	38	8.73
Right TPJ	48	−58	34	6.66
Left Middle Frontal Gyrus	−30	54	4	6.22
Left Inferior Parietal Lobule	−40	−66	42	6.04
Right Middle Frontal Gyrus	56	20	36	4.20
Orbitofrontal cortex	4	50	−18	4.27*
MPFC	−2	52	40	4.13*
Left Middle Temporal Gyrus	−60	−30	−10	3.97
Group > Control				
PC	2	−62	36	7.76
Right TPJ	54	−64	32	5.75
Right Temporal Pole	46	16	−32	5.71
MPFC	−6	54	42	4.85
Right Middle Frontal Gyrus	44	24	28	4.65
Left Inferior Parietal Lobule	−44	−66	42	4.44
MPFC	−10	42	50	4.27
Individual + Group > Control				
PC	0	−60	36	8.45
Right TPJ	48	−58	32	6.32
Left Inferior Parietal Lobule	−42	−66	42	5.60
Left Middle Frontal Gyrus	−32	54	6	5.17
Right Middle Frontal Gyrus	44	24	28	4.94
MPFC	−6	56	44	4.73
Individual > Group				
Right Middle Frontal Gyrus	36	54	6	5.25
Right Posterior Middle Frontal Gyrus	26	12	50	4.87
Left Inferior Parietal Lobule	−46	−56	58	4.32
Group > Individual				
Right Middle Occipital Gyrus	44	−80	14	4.81
Right Fusiform Gyrus	36	−74	−14	4.69
Left Middle Frontal Gyrus	−52	4	40	4.25
Left Posterior Middle Temporal Gyrus	−54	−56	2	4.04

Average peak voxels for regions identified in whole-brain random effects analysis ($p < .001$, $k > 10$ voxels; *= $p < .005$, k > 10 voxels) of the localizer and directed individual vs. group theory-of-mind task in Montreal Neurological Institute (MNI) coordinates. TPJ = temporal parietal junction; PC = precuneus; MPFC = medial prefrontal cortex; STS = superior temporal sulcus.

doi:10.1371/journal.pone.0105341.t001

participants made predictions about what individual or group agents would do in particular situations. In both cases, activation associated with groups was indistinguishable from that associated with consideration of individuals. Whole-brain analyses, conjunction analysis, and ROI analyses all support the conclusion that cognitive processes associated with thinking about the minds of

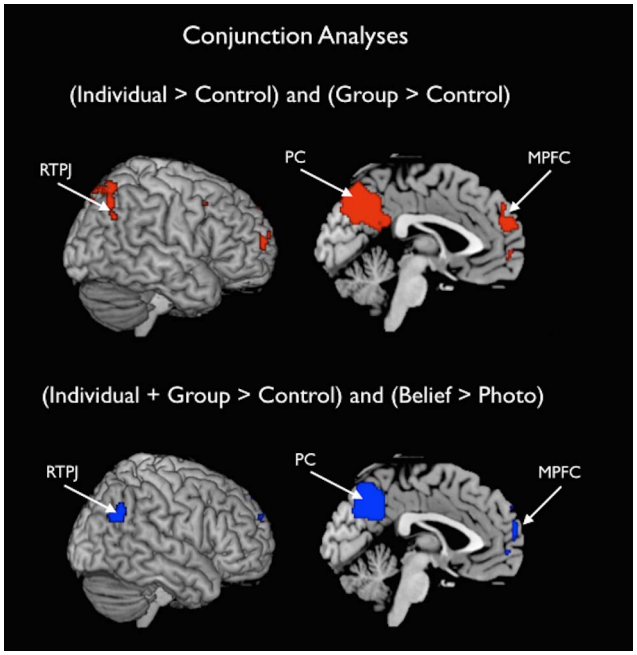

Figure 2. Conjunction analyses. Top: A conjunction analysis revealed conjoint activation in MPFC, TPJ (bilaterally), and precuneus when participants read about the mental states of individuals and groups, compared to a non-mental control condition. Bottom: These regions also overlapped with those recruited by the theory-of-mind localizer. Activations are displayed on a canonical brain image.
doi:10.1371/journal.pone.0105341.g002

individuals were also recruited when participants thought about the 'mind' of a group agent. However, it is worth noting the possibility that participants may have been thinking to some degree about the minds of individual group members, and that this may have accounted for the observed activation in theory-of-mind regions during consideration of group agents. This possibility is weakened, but not completely ruled out, by (a) the fact that, unlike past studies, no individuals were mentioned or shown in the group condition and (b) the observation that perceivers interpret sentences about group mental states as ascribing mental states to the group agent itself in Experiment 1, and (c) the recent observation that the more perceivers think about the 'mind' of the group, the less they think about the minds of its members [8].

Past research has documented the selectivity of the RTPJ for attributing representational mental content, such as beliefs and intentions, to others [22,25,57,61,62], compared to other sorts of attributions, such as those concerning a person's physical appearance, preferences, or personality traits. In this research, neither the mere presence of a person nor the need to make other types of inferences about that person was associated with as much activation in this region as attributing representational mental states. Accordingly, the fact that the RTPJ activated indistinguishably during consideration of individuals and groups (but distinguished both from the inanimate control condition) is an especially compelling suggestion that participants used similar processes for understanding the representational mental states of individuals and group agents.

Although the specific contributions of MPFC to social cognition remain uncertain, this region has been observed to be sensitive to the *target* of mental state ascription. In particular, greater MPFC activation has been associated with interpersonally close others [63–67], and with humanized others [68], compared to those who are more distant or dehumanized. Accordingly, it would not have been surprising to observe reduced MPFC response to group agents compared to individuals. However, the current study observed indistinguishable engagement during consideration of group agents and individuals in a region of MPFC involved in attributing mental states to individuals, as identified by the theory-of-mind localizer, and similar to regions of MPFC associated with mentalizing or theory-of-mind in past studies (according to Neurosynth [70]). Moreover, the individual condition and group condition were associated with greater MPFC activation than the inanimate control condition, suggesting that MPFC's contributions to individual-oriented social cognition are also present during social cognition concerning group agents.

More generally, an abundance of past research has observed greater engagement of brain regions associated with theory-of-mind when perceivers think about certain types of target entities

Table 2. Regions emerging from the conjunction analysis.

Region	X	y	Z
PC	0	−60	36
Right TPJ	48	−60	32
Right Middle Frontal Gyrus	44	24	28
Right Middle Frontal Gyrus	52	16	46
MPFC	12	56	10
MPFC	−6	54	42
Left Middle Frontal Gyrus	−28	52	10
Left Middle Frontal Gyrus	−38	54	−2
Left Middle Frontal Gyrus	−52	20	38
Left Anterior Superior Temporal Gyrus	−34	6	−24
Left TPJ	−52	−66	28
Left Inferior Parietal Lobule	−42	−66	42
Left Middle Temporal Gyrus	−60	−28	−10

Average peak voxels for regions identified in whole-brain conjunction analysis of the individual > control and group > control contrasts ($p < .01$ for each) in Montreal Neurological Institute (MNI) coordinates. TPJ = temporal parietal junction; PC = precuneus; MPFC = medial prefrontal cortex.
doi:10.1371/journal.pone.0105341.t002

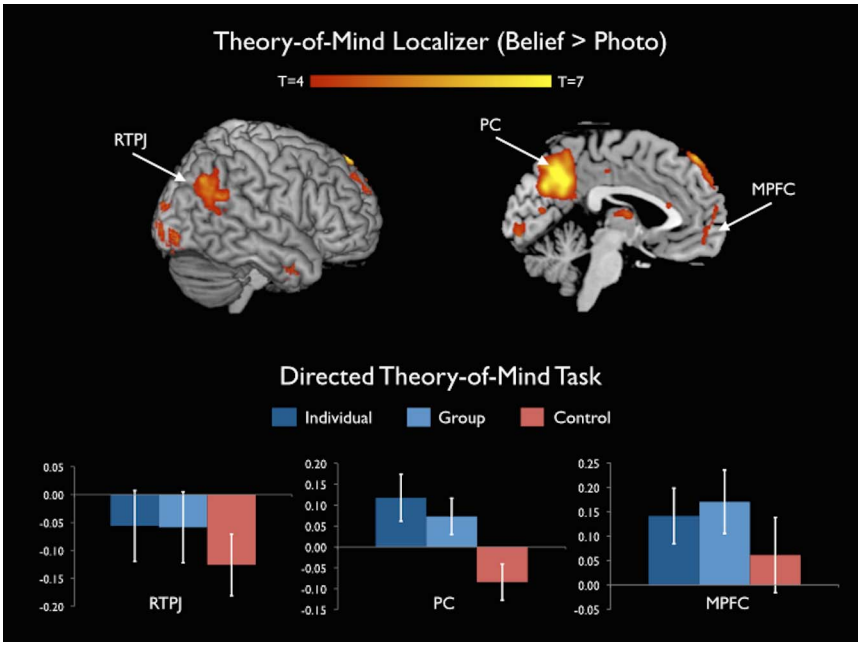

Figure 3. Regions identified by the theory-of-mind localizer. Top: Brain regions emerging from the theory-of-mind localizer (belief > photo; $p < .001$, uncorrected, $k > 10$). Activations are displayed on a canonical brain image. Bottom: Percent signal change (PSC) in BOLD response during the individual, group, and control conditions of the directed theory-of-mind task in regions identified by the independent theory-of-mind localizer.
doi:10.1371/journal.pone.0105341.g003

(humans and, to some degree, other animals) than when they think about other types of target entities (computers, food, furniture); for reviews, see [71–73]. Here, we find just as much activation in brain regions associated with theory-of-mind when people think about group agents as when they think about individual humans, yet a group agent is something very different from a human being or animal, or even from a collection of human beings. Accordingly, the current results are consistent with the possibility that perceivers apply theory-of-mind generally to things that conform to a certain kind of abstract structure [13,74], and that group agents turn out to be among the things that conform to that structure [75]. This possibility draws further support from recent research observing activation in brain regions associated with theory-of-mind during consideration of other non-human agents that display human-like properties [76–78] and is broadly consistent with the observation that brain regions engaged when people construct representations of others' mental states are also engaged when people construct other types of representations that are removed from their current, first-person experience, such as representations of the past or future [79–82].

In sum, people appear in certain respects to treat groups as 'entities' [47]. They assign moral blame to whole organizations as a whole [1], treat whole financial markets as though they have minds of their own [83], and give corporations many of the legal rights enjoyed by individual human beings [4]. In the current studies, we observed that perceivers were willing to attribute mental states to group agents that they did not attribute to the individual members of those groups, and that attributing mental states to group agents was associated with activation in the same brain regions that support ascriptions of mental states to individual people (as confirmed by an independent localizer task). Taken together, these results suggest that in order to understand the striking ways in which people reason about corporations, governments, and other group agents, it may be important to consider the possibility that perceivers sometimes attribute mental states such as beliefs, desires, and intentions not only to the members of such groups but also to the group agent itself.

Supporting Information

Text S1 Stimuli from Experiment 1. Full text of all vignettes and questions.
(PDF)

Text S2 Stimuli from Experiment 2. Full text of the statements and questions from the directed and spontaneous theory-of-mind tasks.
(PDF)

Table S1 Data from Experiment 1. Individual subject responses for each vignette. Condition 1 = 'any member'; 2 = 'each member'; 3 = 'group'.
(PDF)

Table S2 Data from Experiment 2. Mean percent signal change (PSC) for each subject in each condition of the directed and spontaneous theory-of-mind tasks in regions identified by the theory-of-mind localizer.
(PDF)

Acknowledgments

The authors thank Rebecca Cox for assistance with data collection. Imaging data were collected at the Athinoula A. Martinos Center for Biomedical Imaging at MIT.

Author Contributions

Conceived and designed the experiments: ACJ RS JK. Performed the experiments: ACJ DDF. Analyzed the data: ACJ DDF JK. Contributed to the writing of the manuscript: ACJ DDF RS JK.

References

1. Baron J, Ritov I (1993) Intuitions about penalties and compensation in the context of tort law. J Risk Uncertain 7: 17–33.
2. Solan L (2005) Private language, public laws: The central role of legislative intent in statutory interpretation. Georgetown Law J 93.
3. DeMartino B, O'Doherty J, Ray D, Bossaerts P, Camerer C (2013) In the mind of the market: Theory of mind biases value computation during financial bubbles. Neuron 79: 1222–1231.
4. Citizens United v. Federal Election Commission, 558 U.S. 08–205 (2010).
5. Cikara M, Bruneau EG, Saxe RR (2011) Us and Them: Intergroup Failures of Empathy. Curr Dir Psychol Sci 20: 149–153.
6. Cuddy AJC, Fiske ST, Glick P (2008) Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. Adv Exp Soc Psychol 40: 61–149.
7. Haslam N (2006) Dehumanization: an integrative review. Pers Soc Psychol Rev 10: 252–264.
8. Waytz A, Young L (2012) The Group-Member Mind Trade-Off: Attributing Mind to Groups Versus Group Members. Psychol Sci 23: 77–85.
9. Frith U, Frith CD (2003) Development and neurophysiology of mentalizing. Philos Trans R Soc Lond B Biol Sci 358: 459–473.
10. Frith CD, Frith U (2006) The Neural Basis of Mentalizing. Neuron 50: 531–534.
11. Baron-Cohen S, Leslie AM, Frith U (1985) Does the autistic child have a "theory of mind"? Cognition 21: 37–46.
12. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1: 515.
13. Dennett D (1987) True Believers. The Intentional Stance. Cambridge, MAMIT Press13–42.
14. Waytz A, Gray K, Epley N, Wegner DM (2010) Causes and consequences of mind perception. Trends Cogn Sci 14: 383–388.
15. Young L, Camprodon JA, Hauser M, Pascual-Leone A, Saxe R (2010) Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. Proc Natl Acad Sci U S A 107: 6753–6758.
16. Young L, Saxe R (2009) An FMRI investigation of spontaneous mental state inference for moral judgment. J Cogn Neurosci 21: 1396–1405.
17. Goel V, Grafman J, Sadato N, Hallett M (1995) Modeling other minds.
18. Fletcher PC, Happé F, Frith U, Baker SC, Dolan RJ, et al. (1995) Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57: 109–128.
19. 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 U S A 98: 11832–11835.
20. Gallagher HL, Jack AI, Roepstorff A, Frith CD (2002) Imaging the intentional stance in a competitive game. Neuroimage 16: 814–821.
21. Mitchell JP, Heatherton TF, Macrae CN (2002) Distinct neural systems subserve person and object knowledge. 99.
22. Saxe R, Kanwisher N (2003) People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." Neuroimage 19: 1835–1842.
23. Sanfey A, Rilling J, Aronson J, Nystrom L, Cohen J (2003) The Neural Basis of Economic Decision making in the Ultimatum Game. Science (80-) 300: 1755–1758.
24. Mitchell JP, Macrae CN, Banaji MR (2005) Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. Neuroimage 26: 251–257.
25. Saxe R, Powell LJ (2006) It's the thought that counts: Specific brain regions for one component of theory of mind. Psychol Sci 17: 692–699.
26. Jenkins AC, Mitchell JP (2011) Medial prefrontal cortex subserves diverse forms of self-reflection. Soc Neurosci 6: 211–218.
27. Bruneau E, Dufour N, Saxe R (2013) How We Know It Hurts: Item Analysis of Written Narratives Reveals Distinct Neural Responses to Others' Physical Pain and Emotional Suffering. PLoS One 8.
28. Stuss DT, Gallup GG, Alexander MP (2001) The frontal lobes are necessary for "theory of mind".
29. Apperly IA, Samson D, Chiavarino C, Humphreys GW (2004) Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a false-belief task with reduced language and executive demands.
30. Bonnington C (2013) Say goodbye to the iPod classic. Wired. Available: http://www.wired.com/gadgetlab/2013/09/goodbye-ipod-classic/.
31. Branaccio D (2013) Apple wants your old iPhone back. Marketplace.
32. Diallo A (2013) Apple releases iTunes radio, a Pandora alternative. Forbes.
33. Arico A (2010) Folk Psychology, Consciousness, and Context Effects. Rev Philos Psychol 1: 371–393.
34. Arico A, Fiala B, Goldberg RF, Nichols S (2011) The Folk Psychology of Consciousness. Mind Lang 26: 327–352.
35. Kashima Y, Kashima E, Chiu CY, Farsides T, Gelfand M, et al. (2005) Culture, essentialism, and agency: Are individuals universally believed to be more real entities than groups? Eur J Soc Psychol 35: 147–169.
36. Knobe J, Prinz J (2008) Intuitions about consciousness: Experimental studies. Phenomenol Cogn Sci 7: 67–83.
37. Huebner B, Bruno M, Sarkissian H (2010) What Does the Nation of China Think About Phenomenal States? Rev Philos Psychol 1: 225–243.
38. Bloom P, Kelemen D (1995) Syntactic cues in the acquisition of collective nouns. Cognition 56: 1–30.
39. Phelan M, Arico A, Nichols S (2012) Thinking things and feeling things: On an alleged discontinuity in folk metaphysics of mind. Phenomenol Cogn Sci 12: 703–725.
40. Fiske ST, Cuddy a JC, Glick P (2007) Universal dimensions of social cognition: warmth and competence. Trends Cogn Sci 11: 77–83.
41. Blum L (2004) Stereotypes And Stereotyping: A Moral Analysis. Philos Pap 33: 251–289.
42. Steele CM, Spencer SJ, Aronson J (2002) Contending with group image: The psychology of stereotype and identity threat. Advances in Experimental Social PsychologyVol. 34. 379–440.
43. Spears RE, Oakes PJ, Ellemers NE, Haslam SA, editors (1997) The Social Psychology of Stereotyping and Group Life. Blackwell Bublishing.
44. Tajfel H (1982) Social psychology of intergroup relations. Annu Rev Psychol 33: 1–39.
45. Allport GW (1979) The Nature of Prejudice. Basic Books.
46. Bloom P, Veres C (1999) The perceived intentionality of groups. Cognition 71.
47. Hamilton DL, Sherman SJ (1996) Perceiving persons and groups. Psychol Rev 103: 336–355.
48. Contreras JM, Schirmer J, Banaji MR, Mitchell JP (2012) Common Brain Regions with Distinct Patterns of Neural Responses during Mentalizing about Groups and Individuals. J Cogn Neurosci 25: 1406–1417.
49. Abelson RP, Dasgupta N, Park J, Banaji MR (1998) Perceptions of the collective other. Pers Soc Psychol Rev 2: 243–250.
50. Morewedge CK, Chandler JJ, Smith R, Schwarz N, Schooler J (2013) Lost in the crowd: Entitative group membership reduces mind attribution. Conscious Cogn 22: 1195–1205.
51. Aichhorn M, Perner J, Weiss B, Kronbichler M, Staffen W, et al. (2009) Temporo-parietal junction activity in theory-of-mind tasks: falseness, beliefs, or attention. J Cogn Neurosci 21: 1179–1192.
52. Ciaramidaro A, Adenzato M, Enrici I, Erk S, Pia L, et al. (2007) The intentional network: How the brain reads varieties of intentions. Neuropsychologia 45: 3105–3113.
53. Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, Haxby J V (2007) Two takes on the social brain: a comparison of theory of mind tasks. J Cogn Neurosci 19: 1803–1814.
54. Perner J, Aichhorn M, Kronbichler M, Staffen W, Ladurner G (2006) Thinking of mental and other representations: the roles of left and right temporo-parietal junction. Soc Neurosci 1: 245–258.
55. Saxe R, Wexler A (2005) Making sense of another mind: The role of the right temporo-parietal junction. Neuropsychologia 43: 1391–1399.
56. Saxe R (2006) Uniquely human social cognition. Curr Opin Neurobiol 16: 235–239.
57. Jenkins AC, Mitchell JP (2010) Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. Cereb Cortex 20: 404–410.
58. Gallagher HL, Happé F, Brunswick N, Fletcher PC, Frith U, et al. (2000) Reading the mind in cartoons and stories: An fMRI study of "theory of mind" in verbal and nonverbal tasks. Neuropsychologia 38: 11–21.
59. Saxe R (2010) The right temporo-parietal junction: a specific brain region for thinking about thoughts. In: Leslie A, German T, editors. Handbook of Theory of Mind. pp. 1–35.
60. Mitchell JP (2008) Activity in right temporo-parietal junction is not selective for theory-of-mind. Cereb Cortex 18: 262–271.
61. Young L, Dodell-Feder D, Saxe R (2010) What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. Neuropsychologia 48: 2658–2664.
62. Scholz J, Triantafyllou C, Whitfield-Gabrieli S, Brown EN, Saxe R (2009) Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. PLoS One 4.
63. Jenkins AC, Macrae CN, Mitchell JP (2008) Repetition suppression of ventromedial prefrontal activity during judgments of self and others. Proc Natl Acad Sci U S A 105: 4507–4512.
64. Tamir DI, Mitchell JP (2010) Neural correlates of anchoring-and-adjustment during mentalizing. Proc Natl Acad Sci U S A 107: 10827–10832.
65. Ames DL, Jenkins AC, Banaji MR, Mitchell JP (2008) Taking another person's perspective increases self-referential neural processing: Short report. Psychol Sci 19: 642–644.
66. Krienen FM, Tu P-C, Buckner RL (2010) Clan mentality: evidence that the medial prefrontal cortex responds to close others. J Neurosci 30: 13906–13915.
67. Mitchell JP, Macrae CN, Banaji MR (2006) Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others. Neuron 50: 655–663.
68. Harris LT, Fiske ST (2006) Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. Psychol Sci 17: 847–853.
69. Cabeza R, Dolcos F, Graham R, Nyberg L (2002) Similarities and differences in the neural correlates of episodic memory retrieval and working memory. Neuroimage 16: 317–330.

70. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011) Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 8: 665–670.
71. Jenkins AC, Mitchell JP (2011) How has cognitive neuroscience contributed to social psychological theory? In: Todorov A, Fiske ST, Prentice D, editors. Social Neuroscience: Towards Understanding the Underpinnings of the Social MindOxford, UKOxford University Press
72. Van Overwalle F (2009) Social cognition and the brain: A meta-analysis. Hum Brain Mapp30: 829–858.
73. Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7: 268–277.
74. Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7: 268–277.
75. Tollefsen D (2002) Organizations as true believers. J Soc Philos 33: 395–410.
76. Schjoedt U, Stødkilde-Jørgensen H, Geertz AW, Roepstorff A (2009) Highly religious participants recruit areas of social cognition in personal prayer. Soc Cogn Affect Neurosci 4: 199–207.
77. Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, et al. (2008) Can machines think? Interaction and perspective taking with robots investigated via fMRI. PLoS One 3.
78. Castelli F, Happé F, Frith U, Frith C (2000) Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns.
79. Schacter DL, Addis DR, Buckner RL (2007) Remembering the past to imagine the future: the prospective brain. Nat Rev Neurosci 8: 657–661.
80. Buckner RL, Carroll DC (2007) Self-projection and the brain. Trends Cogn Sci 11: 49–57.
81. Addis DR, Wong AT, Schacter DL (2007) Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. Neuropsychologia 45: 1363–1377.
82. Spreng RN, Mar RA, Kim ASN (2009) The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J Cogn Neurosci 21: 489–510.
83. De Martino B, Kumaran D, Seymour B, Dolan RJ (2006) Frames, biases, and rational decision-making in the human brain. Science 313: 684–687.