Spontaneous Mentalizing Predicts the Fundamental Attribution Error

Joseph M. Moran[1,2], Eshin Jolly[1], and Jason P. Mitchell[1]

Abstract

■ When explaining the reasons for others' behavior, perceivers often overemphasize underlying dispositions and personality traits over the power of the situation, a tendency known as the fundamental attribution error. One possibility is that this bias results from the spontaneous processing of others' mental states, such as their momentary feelings or more enduring personality characteristics. Here, we use fMRI to test this hypothesis. Participants read a series of stories that described a target's ambiguous behavior in response to a specific social situation and later judged whether that act was attributable to the target's internal dispositions or to external situational factors. Neural regions consistently associated with mental state inference—especially, the medial pFC—strongly predicted whether participants later made dispositional attributions. These results suggest that the spontaneous engagement of mentalizing may underlie the biased tendency to attribute behavior to dispositional over situational forces. ■

INTRODUCTION

Human beings place a significant premium on understanding why other humans do the things they do. Is the woman sitting next to me at the coffee shop talking so loudly because she is a shrill and high-strung person or simply because she's overcaffeinated? In the absence of additional information about this person (e.g., seeing what she's like before imbibing three espressos), observers might be expected to remain agnostic about whether her behavior reflects stable personality characteristics or more transient features of the situation. However, decades of social psychological research have revealed that perceivers frequently gravitate strongly toward dispositional inferences ("she's a jumpy one!") and inexplicably discount situational influences (the effects of caffeine on behavior; e.g., Jones & Harris, 1967; Kelley, 1967). This psychological bias has been termed the "fundamental attribution error" (Ross, 1977), for the notion that we tend toward making attributions that are fundamental to the person rather than to the situation in which the person finds himself. Influential work reveals that the fundamental attribution error occurs spontaneously when people meet others, without the need for conscious intervention (Winter, Uleman, & Cunniff, 1985). The task for psychologists now is to explain not only the conditions that give rise to the fundamental attribution error (Kelley, 1967) but also the cognitive processes responsible for its use.

A sizeable literature in social cognition reveals that when we aim to understand other people's intentions, we mentalize about them—that is, we seek to represent the contents of others' minds (Frith, Morton, & Leslie, 1991). This feature of human social cognition has also been termed "adopting the intentional stance" (Dennett, 1987), implying the notion that understanding others' actions is best achieved by assuming those actions are guided by intentions. A large number of neuroimaging studies implicate a role for a well-characterized set of brain regions in mentalizing, including medial pFC (MPFC), posterior cingulate cortex, TPJ, and STS (Van Overwalle, 2009; Amodio & Frith, 2006; Mitchell, Banaji, & Macrae, 2005). These regions, among others, have been collectively referred to as "the social brain" (Adolphs, 2003) for their seeming specialization for social knowledge representation. Activation in this network has been observed when we consider others' minds both explicitly (Moran, Lee, & Gabrieli, 2011; Mitchell, Heatherton, & Macrae, 2002) and spontaneously (Ma, Vandekerckhove, Van Overwalle, Seurinck, & Fias, 2010; Moran, Heatherton, & Kelley, 2009) and across a wide range of experimental situations including imputing mind to animated shapes (Martin & Weisberg, 2003), playing cooperative games (McCabe, Houser, Ryan, Smith, & Trouard, 2001), and making moral judgments (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Thus, activation in the social brain regions appears to be associated strongly with our ability to adopt the intentional stance. Recent experimental efforts suggest that the core mentalizing regions (MPFC and TPJ) are activated more than other social brain regions when we infer traits spontaneously (i.e., when task demands encourage different kinds of processing), but that more peripheral regions (such as posterior cingulate cortex and STS) are also brought on-line when we intentionally infer traits

[1]Harvard University, [2]U.S. Army Natick Soldier Research, Development, and Engineering Center

Journal of Cognitive Neuroscience 26:3, pp. 569–576
doi:10.1162/jocn_a_00513

(Ma et al., 2010). Early social neuroscience work also hinted that MPFC in particular is activated in experimental conditions that encourage participants to make dispositional attributions about others' behaviors (Harris, Todorov, & Fiske, 2005), although these activations were not tied to the actual dispositional inferences made.

Here, we test the hypothesis that spontaneous mentalizing may underlie the fundamental attribution error by examining the activation of neural regions involved in mentalizing when participants make attributions about people's behaviors. We tested our hypothesis by having participants undergo fMRI while making attributions about hypothetical behaviors that could have either situational or dispositional causes. We predicted that activation in regions involved in mentalizing would differentiate behaviors about which participants made dispositional attributions versus behaviors they deemed to have a situational cause. Given recent efforts to differentiate the contributions made by different social brain regions (Van Overwalle, 2009), we sought to determine which regions within this network might be differentially involved in making dispositional attributions.

METHODS

Participants and Procedure

Participants were 16 right-handed volunteers (9 men; mean age = 23.0 years, *SEM* = 1.3 years) with no history of psychiatric or neurological disorders. Participants provided informed consent and were compensated in accordance with the regulations of the Committee on the Use of Human Subjects at Harvard University.

We created a set of 48 scenarios about a person's behavior that contained information describing both situational and dispositional causes for that behavior (stories are available at wjh.harvard.edu/~scanlab/Moran_FAE_story_stimuli.pdf). Each scenario had a question associated with it that asked about two possible causes for the person's behavior. One answer implied a situational explanation, whereas the other implied a dispositional explanation. Participants had to choose one of these response options. Our goal was to create scenarios describing actions that were equally attributable to something about the situation or something about the person in the scenario. As such, all scenarios contained mental state information. The fundamental attribution error would thus be revealed in answers that appealed to dispositional causes. It is important to note that such answers are neither correct nor incorrect, as the true cause of the person's behavior is unknown. Scenarios were initially normed by 172 on-line respondents using Amazon's Mechanical Turk service (www.mTurk.com). On-line participants rated the answers on a 7-point scale between the dispositional explanation (1) and the situational explanation (7). In this way, we could determine which scenarios created most consensus among respondents and which did not. During fMRI scanning, participants read the same scenarios with the modification that they made a two-alternative forced-choice response about whether the actor's behavior was more likely attributable to something inherent to her or his dispositions or to aspects of the situation. On-line ratings were correlated with the proportion of situational responses obtained during the neuroimaging experiment, $r(46) = .81, p < .001$. This result demonstrated that ratings obtained during the neuroimaging experiment were representative of ratings for these scenarios obtained outside the MRI scanner environment. Stories were presented for 12 sec, and answer phases were presented for 6 sec. A variable delay (0–6 sec) was interspersed between story and answer phases to allow decomposition of the hemodynamic responses in each phase (Dale & Buckner, 1997). We included a variable interval of 0–8 sec following each trial. Trials were modeled as events with durations for the presentation lengths of the story and answer phases. We also conducted analyses where we modeled answer phases as concluding once participants had made their responses (M = 3.61 sec, *SEM* = 0.1 sec). Activations were broadly identical from both analyses, and we report the results from the analyses modeling the entire presentation of the answer phase.

Stories were conditionalized based on fMRI participants' responses. Specifically, we first computed a measure of "attributional ambiguity" for each story by computing the absolute difference between the number of situational and dispositional attributions it attracted across the participants from the fMRI sample. A score of zero would represent a perfectly ambiguous story for which 50% of participants made a dispositional attribution and 50% made a situational attribution. A score of 100 would imply that the story attracted the same response from all participants. We divided stories into three groups of 16 stories each (high [M = 17], medium [M = 46], and low [M = 80] response ambiguity) and separately modeled responses for the three story types. The rationale for this division was that we felt we were more likely to observe the fundamental attribution error in stories that were relatively ambiguous: If all participants gave a dispositional (or situational) response for a given story, we can infer that the story naturally contains information that leads more directly to one or the other explanation. If, however, a story attracts 50% dispositional and 50% situational responses, we can conclude that there is less trait diagnostic information available in the story and thus that participants who gravitate toward dispositional attributions are committing the fundamental attribution error, as there is no consensus on the true causes of behavior.

To understand how participants viewed these stories, we compared behavioral responses across the story set to determine the proportions of dispositional and situational responses. Overall, participants made more situational ($M_{\text{proportion}} = .63$) than dispositional responses (Wilcoxon signed rank test, $W = 258.5$, $Z = 3.23$, $p < .002$, $r = .47$). This pattern was true for the set of stories

that was low in response ambiguity ($M_{proportion}$ situational = .72; $W = 23$, $Z = 2.33$, $p < .02$, $r = .58$) and was less apparent in both the medium-response ($M_{proportion}$ situational = .64; $W = 31$, $Z = 1.92$, $p = .054$, $r = .48$) and high-response ambiguity story sets ($M_{proportion}$ situational = .54; $W = 31$, $Z = 1.65$, $p > .10$, $r = .41$). Hence, the makeup of these stories led participants to provide situational explanations more often than dispositional explanations. This is not too surprising, given that both the story and answer phases contained information that stated directly the situational pressures on behavior and, in the case of the answer phases, asked participants to consider explicitly whether the situational influence precipitated the behavior. This experimental framing is necessarily unlike real-life episodes that elicit the fundamental attribution error; there, the situational forces are often obscured or only vaguely hinted at, and we are not asked to weigh the relative merits of situational and dispositional forces. Thus, even when the situational forces are made explicit because of experimental constraints, there are still significant numbers of stories for which participants opt for the dispositional explanation, and in the case of the high-response ambiguity stories, this proportion was not different from the proportion of stories that attracted situational responses.

Imaging Acquisition and Analysis

Functional data were acquired using a gradient-echo echo-planar pulse sequence (repetition time = 2 sec, echo time = 35 msec, flip angle $\alpha = 90°$) on a 3T Siemens Tim Trio MRI scanner (Siemens, Erlangen, Germany). Images were acquired using 36 axial, interleaved slices with a thickness of 3 mm (0.54 mm skip) and 3 × 3 mm in-plane resolution. Functional images were preprocessed and analyzed using SPM (Wellcome Department of Cognitive Neurology, London, United Kingdom) and custom software (spm8w, Kelley and Heatherton Labs, Dartmouth College, Hanover, NH). Data were realigned within and across runs to correct for head movement, unwarped to correct for geometric distortions, and transformed into a standard anatomical space (2 mm isotropic voxels) based on the ICBM-152 brain template (Montreal Neurological Institute). Normalized data were then spatially smoothed (8 mm FWHM) using a Gaussian kernel. Finally, using custom artifact detection software (www.nitrc.org/projects/artifact_detect), individual runs were analyzed on a participant-by-participant basis to find outlier time points. Specifically, we excluded volumes during which participant head motion exceeded 0.5 mm or 1° and volumes in which the overall signal for that time point fell more than three standard deviations outside the mean global signal for the entire run. The design matrix included regressors for dispositional and situational responses across story and answer phases for each of the sets of high-, medium-, and low-response ambiguity stories separately (for a total of 12 experimental conditions). The trials were modeled using a canonical hemodynamic response function and covariates of no interest (session mean and linear trend, outlier time points excluded as above). In cases where participants did not respond, story and answer phases for those trials were entered into "no response" conditions that were ignored in further analyses ($M = 4.0\%$, $SEM = 1.2\%$). Analysis was performed individually for each participant, and contrast images were subsequently entered into second-level analyses treating participants as a random effect.

Two whole-brain analyses were performed. In our first analysis, we identified brain regions that responded more when participants made dispositional versus situational attributions across all story types and phases (i.e., across low-, medium- and high-ambiguity stories and across background and answer phases). This analysis allowed us to make a general determination about the brain regions associated with dispositional attributions (regardless of whether there was consensus across participants about the sources of actors' behaviors in the stories). In our second analysis, we narrowed down our stimulus set to just those stories that were maximally ambiguous across our group of participants. This analysis would control for the content of the story phases, as identical scenarios would appear in the dispositional and situational conditions across participants. Because we aimed to determine whether activation in mentalizing brain regions predict how participants will respond, in this analysis, we also considered brain activations in just the story segments of each scenario. Thus, this analysis compared subsequently dispositional stories to subsequently situational stories. Peak coordinates in our whole-brain analyses were identified at the group level using a statistical criterion of 56 or more contiguous voxels at a voxel-wise threshold of $p < .005$. A Monte Carlo simulation implemented in MATLAB determined that these thresholds corresponded with a corrected voxel-wise threshold of $p < .05$ (Slotnick, Moo, Segal, & Hart, 2003). In the simulation, random noise was created with the same voxel dimensions as our preprocessed data (and smoothed with the same 8 mm kernel), with the constraint that the proportion of falsely active voxels was 0.005. After 1000 simulations, the probability of each cluster size was determined, and the cluster extent that yielded $p < .05$ (56 voxels) was selected for use in voxel extent thresholding.

RESULTS

Our initial whole-brain analysis collapsed across both ambiguity and phase to compare all dispositional versus all situational stories. This analysis allowed us to determine the brain regions involved in making dispositional responses across ambiguous and unambiguous stories and across story and answer phases. These brain regions would be on average more active across all phases and story kinds for dispositional versus situational responses. Brain regions showing greater activation for dispositional > situational attributions included, from

anterior to posterior, bilateral dorsomedial pFC (dmPFC) and superior frontal gyrus, ACC, left anterior insula, right inferior frontal gyrus, right posterior STS, right TPJ, and precuneus (Figure 1A and Table 1A). No regions were more active for situational versus dispositional attributions in the reverse contrast. The results of these analyses imply that regions involved in orienting to others' mental states are also engaged when we attribute the causes of someone's behavior to their internal dispositions to act in such a way.

However, this analysis did not control for the stories' content: An alternative possibility is that "dispositional" stories simply contained more mental state information than did "situational" stories. Because we collapsed across both phases, a second possibility is that greater activations in mentalizing regions during dispositional stories were driven by participants paying attention just to the mental state information provided in the dispositional answer they chose during the answer phase. Greater activation in regions responsible for representing mental state information in either circumstance would thus be unsurprising, because they would simply be responding to the presence of mental state information. To control for these confounds, we performed a second analysis, which limited its focus to just the set of 16 scenarios that attracted approximately even numbers of dispositional (46%) and situational responses (54%, recall that these proportions were not significantly different from one another). This analytical choice ensured that across our participants (i.e., at the level of random effects) those scenarios appearing in the dispositional condition were the same as those scenarios appearing in the situational condition. Any brain responses in mentalizing regions would therefore not have occurred as a result of the presence of mental state information alone. In addition, we restricted this

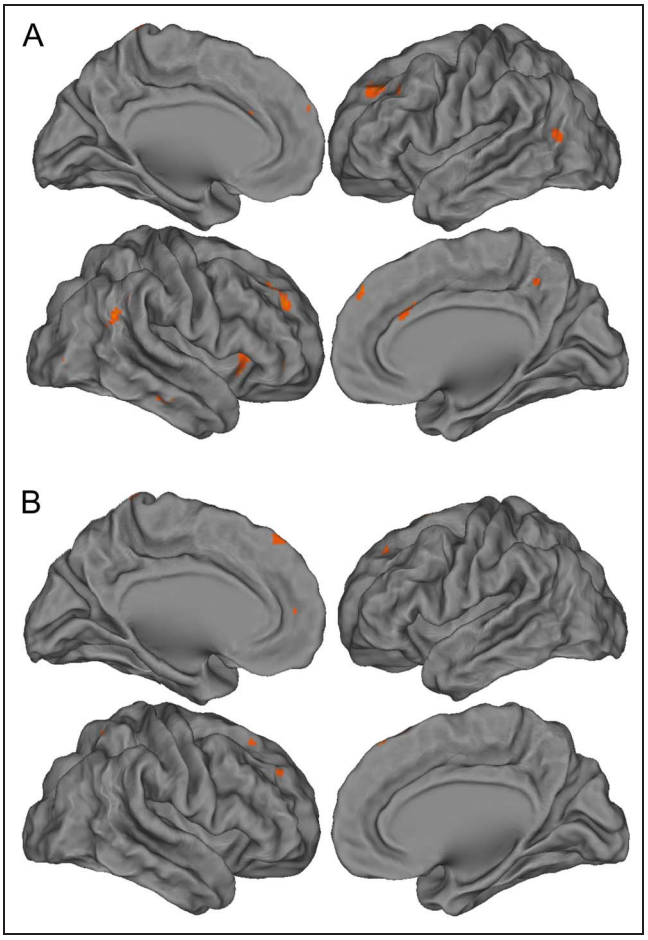

Figure 1. (A) Regions showing greater activations for all dispositional > situational attributions, collapsed across all stories and both story and answer phases. There was greater activation in dmPFC, superior frontal gyrus, ACC, right inferior frontal gyrus, right middle temporal gyrus, the precuneus, and the right TPJ. (B) Regions predicting dispositional attributions: There was greater activation in dmPFC, superior frontal gyrus, and the STS during subsequently situational > subsequently dispositional stories. These regions, implicated in mentalizing, predicted which participants would commit the fundamental attribution error when reading about an actor's behavior.

Table 1. All Scenarios: Peak Voxel and Cluster Size for All Regions Obtained from the Contrast of Dispositional > Situational Attributions across Both Story and Answer Phases

Anatomic Label	*x*	*y*	*z*	*Peak t*	*Voxel Extent*
A. Dispositional > Situational					
(r) Superior frontal gyrus	18	49	32	4.18	260
(r) dmPFC	10	47	44	3.91	
	6	55	42	3.69	
(l) dmPFC	−12	45	44	3.25	
(r) Inferior frontal gyrus	42	41	40	4.62	79
(l) Superior frontal gyrus	−22	33	54	3.64	109
ACC	8	27	32	3.91	91
(r) Inferior frontal gyrus	48	25	0	4.51	352
	40	25	4	3.92	
(l) Anterior insula	−24	15	12	5.86	148
(r) STS	62	−31	−10	4.41	216
	54	−35	−6	4.15	
	62	−21	−22	3.40	
(r) TPJ	58	−49	26	4.01	160
	56	−59	20	3.82	
	48	−51	12	3.50	
Precuneus	12	−49	48	3.46	70
	2	−45	46	3.29	
B. Situational > Dispositional					
No significant differences					

analysis to just the story phases. Analysis was restricted to the story phase because these activations would occur before participants read the answer phases, and hence they would not be attributable to the participants considering only the answer that implied a disposition and ignoring the situational answer.

In this analysis, for each participant, the story phase of each ambiguous story was conditionalized as a function of whether that participant subsequently attributed the actor's behavior to internal dispositions or to external situational factors. As such, this analysis identified neural responses that predict whether a perceiver would attribute dispositional or situational causes to an actor's behavior. A whole-brain, random-effects analysis of subsequently dispositional stories > subsequently situational stories revealed greater activation in a few regions: dmPFC, superior frontal gyrus, aCC, and posterior middle temporal gyrus (Figure 1B and Table 2A). Each of these regions was implicated in the initial analysis comparing all dispositional versus all situational stories. Thus, it appears a subset of the mentalizing network predicts which participants will later make dispositional attributions about actors' behaviors. To the extent that these regions have been linked consistently to mentalizing about other minds (Amodio & Frith, 2006; Frith & Frith, 1999), these results suggest that perceivers were more likely to make a dispositional attribution about a person's actions to the extent that they spontaneously attended to the actor's internal mental states. Regions that were more active for subsequently situational stories > subsequently dispositional stories were the left temporal pole and left amygdala (Table 2B).

Table 2. Ambiguous Scenarios: Peak Voxel and Cluster Size for All Regions Obtained from (a) the Contrast of Subsequently Dispositional Stories > Subsequently Situational Stories and (b) the Contrast of Subsequently Situational Stories > Subsequently Dispositional Stories

Anatomic Label	*x*	*y*	*z*	*Peak t*	*Voxel Extent*
A. Dispositional > Situational					
(l) dmPFC	−10	41	42	3.74	148
	−6	45	42	3.41	
	−12	35	56	3.31	
(r) dmPFC	10	41	54	4.05	67
	14	31	60	3.48	
Superior frontal gyrus	20	33	54	3.25	
ACC	−18	25	28	4.10	83
(r) Middle temporal gyrus	58	−15	−18	4.88	66
	46	−21	−20	3.92	
B. Situational > Dispositional					
(l) Temporal pole	−30	11	−22	4.69	226
	−46	11	−26	4.26	
(l) Amygdala	−28	−1	−24	3.73	

DISCUSSION

We found that the activation of a particular subset of brain regions implicated in mental state understanding—including MPFC and posterior lateral temporal cortices—predicted whether perceivers attributed dispositional or situational causes to another person's ambiguous behavior. In addition, because these activations occurred before participants read the answer phases, they could not be attributed to the participants considering only the answer that implied a disposition and ignoring the situational answer. In particular, MPFC has been consistently associated with forming impressions about and representing the minds of others across multiple experimental paradigms, such as considering others' character traits (Moran et al., 2011) and making moral judgments (Greene et al., 2001). Furthermore, recent evidence suggests that regions of MPFC more ventral to those observed here are used to represent trait knowledge about others (Ma et al., in press). Spontaneous MPFC activation may thus reflect the involvement of cognitive mechanisms designed to extract trait knowledge from exemplars of behavior. As such, these findings suggest that the human tendency to attribute others' actions to dispositional causes—the fundamental attribution error—may depend on the spontaneous activation of these regions in social situations.

Interestingly, in a powerful automated parcellation that converged on a solution separating the brain into seven major networks, each of these regions was implicated as part of the same grouping—the default mode network (Yeo et al., 2011). This fact suggests an intriguing possibility: If these regions are marked by high resting metabolic activity and their activity is associated with dispositional attributions about others' behavior, then it may be unsurprising that perceivers naturally default to making such dispositional (rather than situational) attributions. One of the major puzzles for social psychologists has been explaining why people are so prone to committing the fundamental attribution error, even when the situational influences are made obvious (Gilbert & Malone, 1995). Perhaps a lifetime of reflexively considering others' mental states and intentions creates a default strategy for understanding other minds through reference to their dispositional characteristics, one that might require significant cognitive effort to overcome (Gilbert, Pelham, & Krull, 1988). In other words, one of the consequences of the link between the default network and dispositional attributions is that perceivers may default to using the intentional stance when it is not useful—the fundamental attribution error. There are at least two directly testable predictions from this view. First, individuals with reduced default mode network activity, such as those with autism

(Kennedy, Redcay, & Courchesne, 2006), should be less likely to commit the fundamental attribution error. Although Autism Spectrum Disorder (ASD) individuals can successfully infer traits from behavior (Ramachandran, Mitchell, & Ropar, 2009), the possibility that neurotypicals spontaneously do this more often than ASD individuals awaits further investigation. Second, using techniques gained from real-time fMRI where one can measure ongoing activity levels in circumscribed brain regions (deCharms et al., 2004), it would be possible to create two conditions by presenting the story segments during periods in which the activity of these regions is alternately high and low and to determine whether participants are more likely to commit the fundamental attribution error when naturally fluctuating activity in these regions is high. This result would provide converging evidence for the hypothesis that high resting activity in the default mode network leads to the unwarranted inference of dispositions.

An alternative possible explanation is that activation in the default mode regions more generally (and specifically in the context of our task) reflects the default mode network's role in abstraction and in taking a more distal perspective. Support for this idea comes from a study showing that higher construal (i.e., more abstract processing) of both social and nonsocial items produced activation in dmPFC (Baetens, Ma, Steen, & Van Overwalle, in press). Although in our study the dispositional answers certainly require greater abstraction than our situational answers (e.g., "Daniel returns the money because he is honest" vs. "…because other people were watching him"), two facts caution against this interpretation. First, that medial prefrontal cortical activations predict later dispositional responses before the response options are presented suggests that the specific process of abstraction encouraged by choosing among the alternative answers is not responsible for the observed activations. Second, recent data from our laboratory (Tamir & Mitchell, 2011) argue just the opposite interpretation. In that article, default mode network regions (including MPFC) were more active in a series of proximal versus distal judgments in a number of different domains (spatial, temporal, social, and hypothetical [self vs. a hypothetical self]). These results argue against the interpretation that greater default mode network activation might simply reflect greater abstraction or processing of a distant perspective and are contrasted with the Baetens et al. (in press) finding in that they manipulate distance as opposed to simply the level of construal (or level of processing) required for task completion.

Other regions differentiated dispositional from situational attributions, but only when we collapsed across all stories and across both story and answer phases. These regions included the precuneus, TPJ, anterior insula, and right inferior frontal gyrus. Although each of these regions has been implicated in social cognition research (Van Overwalle, 2009; Adolphs, 2003; Saxe & Kanwisher, 2003; Fletcher et al., 1995), their involvement in the current study was nevertheless not predictive of later dispositional attributions. Thus, spontaneous activation of the MPFC, ACC, and lateral temporal cortices may be indicative of spontaneous mental state representation, whereas activation in the larger social brain network may be indicative of more direct representation of mental states. These findings fit nicely with findings from the literature on trait inferences, in which core mentalizing regions are recruited during both spontaneous and intentional trait inferences, whereas extended mentalizing regions are recruited only during intentional trait inferences (Ma et al., 2010). The sole difference between findings is that the TPJ was recruited for spontaneous trait inferences in Ma et al. (2010) but was not predictive of dispositional inferences in the current findings. One possibility, albeit conjectural, is that the stimuli eliciting spontaneous trait inferences in Ma et al. (2010) may have contained more information pertaining to temporary goals or intentions (Van Overwalle, 2009), whereas this was not the case for the ambiguous dispositional items in the present work.

Other recent work from the same group (Kestemont, Vandekerckhove, Ma, Van Hoeck, & Van Overwalle, 2013; Ma, Vandekerckhove, Van Hoeck, & Van Overwalle, 2012) has sought to distinguish between person and situation attributions. In their work, Kestemont et al. observed greater activation in mentalizing regions (e.g., MPFC and TPJ) for both situational and person (dispositional) attributions relative to a nonsocial control. Although mentalizing activations during situational attributions were not observed in our data, a crucial difference between studies is that we did not use a nonsocial control condition. To the degree that situation attributions in Kestemont et al.'s (2013) study contained person information, one would expect greater activation in regions such as MPFC and TPJ for a social (situation attributions) versus nonsocial (semantic truth judgments) task.

Because participants regularly make situational attributions about their own behavior (Jones & Nisbett, 1971), it is possible that taking another's perspective would lead to greater consideration of the situational forces guiding that person's behavior. MPFC has been regularly implicated in taking the perspective of another individual (D'Argembeau et al., 2007), and so we might have expected this region's involvement when participants made situational rather than dispositional attributions. Because this pattern did not emerge in the present data, we can be somewhat confident that perspective-taking did not play a part in participants' judgments, at least in the context of the present stories. Future investigation could directly manipulate participants' requirements to take others' perspective to understand the role of this cognitive process in making attributions about people's behavior.

Rather than activations in MPFC for predicting situational attributions, we instead saw greater activation in a single region in the left anterior temporal lobe (ATL) that extended into the amygdala. Although a significant body of work suggests a role for the ATL in social cognition in

general (Olson, Plotzker, & Ezzyat, 2007), its role appears to be more circumscribed to the acquisition of social semantic knowledge (Simmons, Reddish, Bellgowan, & Martin, 2010) rather than in mental state representation. A recent review suggests that this region is specialized for the binding of complex perceptual inputs (like the sights and sounds implied by our scenarios) to visceral emotional responses (which fits with the locus of activation in the amygdala observed here; Olson et al., 2007). Although we did not have clear hypotheses about regions predicting situational attributions, we do note that identical social information was present in both conditions, thus suggesting that something about the activations in the ATL and amygdala represent a cognitive process that led participants to greater use of the situational explanation in those trials. The possibility of these regions signaling the salience of situational causes of behavior is an intriguing one that awaits further investigation.

Finally, we note that both the "situational" and "dispositional" options include information about the target's intentions. Theorists in the attribution literature have long noted the apparent misnomer of "situational causations" in fact occurring because of the actor's mental state and not because of the situation per se. Gilbert (1998) discusses behaviors that attract situational explanations (e.g., "He ran away because there was a snake.") that are in reality caused by "ordinary dispositions" (e.g., "I ran away from the rattlesnake because I dislike being injected with venom."). These are contrasted with behaviors that occur because of "extraordinary dispositions" (e.g., "He stayed to tackle the rattlesnake because he is foolhardy" [Gilbert, 1998]). Trope (1989) similarly distinguishes between two kinds of inference regarding dispositions; one leading to behavior-specific causes (like Gilbert's "ordinary" dispositions), and the other leading to general (or "extraordinary dispositional") causes. Along those lines, we aimed to distinguish between the responses available to participants for our stories such that the "dispositional" answers we made available were more distal, general, or truly dispositional, whereas the "situational" answers, which nevertheless reflected actors' intentions, were more proximal, behavior-specific, and reflected normative responses to the situational forces described in the stories. That being said, we feel that the crucial analysis in this article (the prediction of dispositional responses by dmPFC activation) does not hinge upon whether or not the two answers both involve the inference of intentions. We analyzed data from the stories preceding the participants' responses and thus from the period before participants were aware of the available dispositional and situational options for a given story. This argues that the brain responses in dmPFC that predict subsequent dispositional answers reflect spontaneous inferences of extraordinary traits and dispositions, regardless of whether the dispositional and situational answers both require the representation of mental states.

In conclusion, we suggest that mentalizing—the consideration of others' mental states—may be the very reason we make errors in attribution. Our default tendency to adopt the intentional stance, as characterized by recruitment of the core social brain regions, seems a wonderfully adaptive strategy that nevertheless comes with significant costs: biasing our inferential ability in favor of dispositional explanations that are not always warranted.

Acknowledgments

This work was supported by a seed grant from Harvard University to J. P. M. The authors thank Diana Bartenstein and T. J. Eisenstein for assistance.

Reprint requests should be sent to Dr. Joseph M. Moran, 52 Oxford St., 290.01, Center for Brain Science, Harvard University, Cambridge, MA 02138, or via e-mail: jmoran@wjh.harvard.edu.

REFERENCES

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience, 4,* 165–178.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7,* 268–277.

Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (in press). Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience*. doi: 10.1093/scan/nst048.

Dale, A. M., & Buckner, R. L. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping, 5,* 329–340.

D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., et al. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Cognitive Neuroscience, 19,* 935–944.

deCharms, R. C., Christoff, K., Glover, G. H., Pauly, J. M., Whitfield, S., & Gabrieli, J. D. (2004). Learned regulation of spatially localized brain activation using real-time fMRI. *Neuroimage, 21,* 436–443.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

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

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

Frith, U., Morton, J., & Leslie, A. M. (1991). The cognitive basis of a biological disorder—Autism. *Trends in Neurosciences, 14,* 433–438.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 2, pp. 173–220). New York: Oxford University Press.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117,* 21–38.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness—When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54,* 733–740.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105–2108.

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: Neuro-imaging dispositional inferences, beyond theory of mind. *Neuroimage, 28,* 763–769.

Jones, E. E., & Harris, V. A. (1967). Attribution of attitudes. *Journal of Experimental Social Psychology, 3,* 1–24.
Jones, E. E., & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behavior*. Morristown, NJ: General Learning Press.
Kelley, H. H. (1967). *Attribution theory in social psychology*. Paper presented at the Nebraska Symposium on Motivation.
Kennedy, D. P., Redcay, E., & Courchesne, E. (2006). Failing to deactivate: Resting functional abnormalities in autism. *Proceedings of the National Academy of Sciences, U.S.A., 103,* 8275–8280.
Kestemont, J., Vandekerckhove, M., Ma, N., Van Hoeck, N., & Van Overwalle, F. (2013). Situation and person attributions under spontaneous and intentional instructions: An fMRI study. *Social Cognitive and Affective Neuroscience, 8,* 481–493.
Ma, N., Baetens, K., Vandekerckhove, M., Kestemont, J., Fias, W., & Van Overwalle, F. (in press). Traits are represented in the medial prefrontal cortex: An fMRI adaptation study. *Social Cognitive and Affective Neuroscience*. doi: 10.1093/scan/nst098.
Ma, N., Vandekerckhove, M., Van Hoeck, N., & Van Overwalle, F. (2012). Distinct recruitment of temporo-parietal junction and medial prefrontal cortex in behavior understanding and trait identification. *Social Neuroscience, 7,* 591–605.
Ma, N., Vandekerckhove, M., Van Overwalle, F., Seurinck, R., & Fias, W. (2010). Spontaneous and intentional trait inferences recruit a common mentalizing network to a different degree: Spontaneous inferences activate only its core areas. *Social Neuroscience, 6,* 123–138.
Martin, A., & Weisberg, J. (2003). Neural foundations for understanding social and mechanical concepts. *Cognitive Neuropsychology, 20,* 575–587.
McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences, U.S.A., 98,* 11832–11835.
Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *Neuroimage, 28,* 757–762.
Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences, U.S.A., 99,* 15238–15243.
Moran, J. M., Heatherton, T. F., & Kelley, W. M. (2009). Modulation of cortical midline structures by implicit and explicit self-relevance evaluation. *Social Neuroscience, 4,* 197–211.
Moran, J. M., Lee, S. M., & Gabrieli, J. D. (2011). Dissociable neural systems supporting knowledge about human character and appearance in ourselves and others. *Journal of Cognitive Neuroscience, 23,* 2222–2230.
Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: A review of findings on social and emotional processing. *Brain, 130,* 1718–1731.
Ramachandran, R., Mitchell, P., & Ropar, D. (2009). Do individuals with autism spectrum disorders infer traits from behavior? *Journal of Child Psychology and Psychiatry, 50,* 871–878.
Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 173–220). New York: Academic Press.
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.
Simmons, W. K., Reddish, M., Bellgowan, P. S., & Martin, A. (2010). The selectivity and functional connectivity of the anterior temporal lobes. *Cerebral Cortex, 20,* 813–825.
Slotnick, S. D., Moo, L. R., Segal, J. B., & Hart, J., Jr. (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. *Brain Research: Cognitive Brain Research, 17,* 75–82.
Tamir, D. I., & Mitchell, J. P. (2011). The default network distinguishes construals of proximal versus distal events. *Journal of Cognitive Neuroscience, 23,* 2945–2955.
Trope, Y. (1989). Levels of inference in dispositional judgment. *Social Cognition, 7,* 296–314.
Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30,* 829–858.
Winter, L., Uleman, J. S., & Cunniff, C. (1985). How automatic are social judgments. *Journal of Personality and Social Psychology, 49,* 904–917.
Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology, 106,* 1125–1165.