An fMRI Investigation of Spontaneous Mental State Inference for Moral Judgment

Liane Young[1,2] and Rebecca Saxe[2]

Abstract

■ Human moral judgment depends critically on ''theory of mind,'' the capacity to represent the mental states of agents. Recent studies suggest that the right TPJ (RTPJ) and, to lesser extent, the left TPJ (LTPJ), the precuneus (PC), and the medial pFC (MPFC) are robustly recruited when participants read explicit statements of an agent's beliefs and then judge the moral status of the agent's action. Real-world interactions, by contrast, often require social partners to infer each other's mental states. The current study uses fMRI to probe the role of these brain regions in supporting spontaneous mental state inference in the service of moral judgment. Participants read descriptions of a protagonist's action and then either (i) ''moral'' facts about the action's effect on another person or (ii) ''nonmoral'' facts about the situation. The RTPJ, PC, and MPFC were recruited selectively for moral over nonmoral facts, suggesting that processing moral stimuli elicits spontaneous mental state inference. In a second experiment, participants read the same scenarios, but explicit statements of belief preceded the facts: Protagonists believed their actions would cause harm or not. The response in the RTPJ, PC, and LTPJ was again higher for moral facts but also distinguished between neutral and negative outcomes. Together, the results illuminate two aspects of theory of mind in moral judgment: (1) spontaneous belief inference and (2) stimulus-driven belief integration. ■

INTRODUCTION

When evaluating the moral status of an action, we consider not only its consequences but also the beliefs and intentions of the agent (Cushman, in press; Mikhail, 2007; Young, Cushman, Hauser, & Saxe, 2007; Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006; Cushman, Young, & Hauser, 2006; Knobe, 2005; Baird & Moses, 2001; Zelazo, Helwig, & Lau, 1996). Consider a daycare worker who serves spoiled meat to the children in his care. Typically, our moral judgment of the worker depends on, among other factors, his mental state at the time of the action. If he believed that the meat was fresh and safe to eat because of its sealed packaging and expiration date, we might exculpate him for his tragic but innocent error. If, however, he believed that he was making the children sick, we would regard the same action as morally blameworthy. Consistent with these intuitions, prior behavioral experiments suggest that when the agent's beliefs (or intentions) are explicitly stated in the scenario, participants' moral judgments depend significantly more on this mental state information than on the actual outcome of the action (for a review, see Baird & Astington, 2004).

The cognitive ability to think about another person's beliefs and intentions is known as ''theory of mind.'' In fMRI studies, a consistent group of brain regions is recruited when participants view story and cartoon stimuli that depict a character's mental state (in nonmoral contexts): the medial pFC (MPFC), the right TPJ (RTPJ) and the left TPJ (LTPJ), and the precuneus (PC) (Ruby & Decety, 2003; Saxe & Kanwisher, 2003; Vogeley et al., 2001; Gallagher et al., 2000; Fletcher et al., 1995). Of these regions, the RTPJ appears to be particularly selective for processing mental states with representational content (e.g., beliefs; Aichorn, Perner, Kronbichler, Staffen, & Ladurner, in press; Ciaramidaro et al., 2007; Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007; Perner, Aichorn, Kronbichler, Staffen, & Ladurner, 2006; Saxe & Wexler, 2005). For example, the RTPJ response is high when participants read stories that describe a character's beliefs but low during stories containing other socially relevant information about a character, including physical appearance, cultural background, or even internal subjective sensations such as hunger or fatigue (Saxe & Powell, 2006).

The precise role of these brain regions in theory of mind for moral judgment has been the topic of recent research (Young & Saxe, 2008; Young et al., 2007). The same regions that support theory of mind in nonmoral contexts are robustly recruited when participants read explicit statements of an actor's beliefs (e.g., ''The daycare worker believed that the meat was unsafe to eat'') and then judge the moral status of the action (e.g., ''The daycare worker served the meat to the children''). The findings further suggest a distinction between two

[1]Harvard University, [2]Massachusetts Institute of Technology

separate component processes of theory of mind for moral judgment: belief encoding and belief integration (Young & Saxe, 2008). Encoding consists of forming an initial representation of the protagonist's belief, as stated in the stimuli. Brain regions for encoding are recruited when belief information is first presented. Integration, by contrast, consists of using of the belief in flexible combination with the outcome for constructing a moral judgment. The integration response should therefore be modulated by the outcome.

The focus of the current study is on a third possible aspect of theory of mind for moral judgment: spontaneous belief inference, when explicit beliefs are unavailable. Indeed, in real life, unlike experimental scenarios, direct access to the contents of other minds is very rare. Usually, observers are presented with just the outcome of the action—the sick children, in the case of the daycare worker. How does moral judgment proceed when information about the agent's mental state is not explicitly available? One possibility is that observers simply rely on observable facts and condemn the daycare worker based on the outcome alone. An alternative view is that moral judgment obligatorily involves mental state considerations; if these features of an action are not explicitly provided, observers will spontaneously attempt to establish them.

We therefore hypothesized that *morally relevant facts* about an action—in particular, facts about the action's effects on other people, deleterious or not—prompt spontaneous belief inference. In other words, reading such facts should result in spontaneous consideration of the actor's mental state—in the absence of any direct reference to the actor's mind. By contrast, other facts elaborating on the situation should not induce people to consider the actor's mental state. In Experiment 1, we tested these hypotheses, using as a dependent measure the time course and magnitude of activation in the brain regions previously implicated in reasoning about mental states. Participants read short verbal descriptions of an agent's action (e.g., ''The daycare worker serves the meat to the children''). Then, participants read a fact about the situation that was either morally relevant (e.g., ''The meat was actually unsafe to eat'') or morally irrelevant (e.g., ''The meat was purchased at the local market''). Half of the moral facts described negative outcomes, and half described neutral outcomes. Although neither the moral nor the nonmoral facts included any reference to the agent's mental state, we hypothesized that participants would spontaneously consider the agent's mental state when reading the moral, but not nonmoral, facts. As a consequence, we predicted that brain regions in the theory of mind network would show enhanced activity selectively for presentation of the moral facts. (We note the possibility that this enhanced activity could reflect increased consideration of the potential *victim's* mental state, although in our previous research it is the direct *manipulation of the actor's belief* as described explicitly in the moral scenarios that modulates the RTPJ response. We have not, however, manipulated the victim's belief; that the RTPJ response may reflect representation of the victim's belief thus deserves further empirical attention.) A further question was whether the effect would be greater for negative outcomes than neutral outcomes, or whether instead information about *any* effect on other people would elicit enhanced activation in this context.

We note the current study focuses on beliefs in light of previous fMRI studies implicating the RTPJ in processing beliefs (e.g., Saxe & Kanwisher, 2003) and behavioral studies, suggesting a specific role for beliefs in moral judgment (Cushman, in press). We recognize that other mental state factors (e.g., intentions) also contribute to mature moral judgment and provide a rich topic for future research (e.g., Mikhail, 2007; Borg et al., 2006; Cushman et al., 2006; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Given previous research indicating a particularly selective role for the RTPJ both in processing explicit beliefs for predicting and explaining behavior in nonmoral contexts (e.g., Perner, Aichorn, Kronbichler, Wolfgang, & Laddurner, 2006; Saxe & Powell, 2006) and in encoding and integrating explicit beliefs for moral judgment (Young & Saxe, 2008; Young et al., 2007), we predicted the most robust and selective activation patterns for the RTPJ in this experiment. Other regions, including the MPFC, have been implicated more broadly in social and moral cognition (e.g., Mitchell, Macrae, & Banaji, 2006; Saxe & Powell, 2006; Greene et al., 2001, 2004; Adolphs, 2003).

In Experiment 2, we sought to further characterize the interaction between theory of mind and moral judgment when explicit belief information was provided before any outcome information. As discussed above, our previous results suggest that theory of mind brain regions are recruited for integrating explicit beliefs with outcomes when both factors are present. We therefore predicted the same integration pattern, reflecting the influence of outcome, in Experiment 2. We predicted that the response, at the time of the outcome, would be influenced by whether explicit beliefs were previously presented (requiring integration) or not (requiring spontaneous inference).

EXPERIMENT 1

Methods

Fourteen naive right-handed participants (Harvard College undergraduates, aged 18–22 years, eight women) participated in the study for payment. All participants were native English speakers, had normal or corrected-to-normal vision, and gave written informed consent in accordance with the requirements of internal review

board at MIT. Participants were scanned at 3T (at the MIT scanning facility in Cambridge, MA) using twenty-six 4-mm-thick near-axial slices covering the whole brain. Standard EPI procedures were used (TR = 1.5 sec, TE = 40 msec, flip angle 90°).

Stimuli consisted of three variations of 48 scenarios for a total of 144 stories (for a sample scenario, see Figure 1; full text available for download at http://www.mit.edu/~lyoung/files/). Participants read descriptions of a protagonist's action and then either (i) ''moral'' facts about the action's effect (harmful or harmless) on another person or (ii) ''nonmoral'' facts about the situation. Stories were presented in cumulative segments:

- Background: the protagonist's action in context (identical across conditions).
- Fact: *moral* facts about the action's outcome (negative or neutral) or *nonmoral* facts about the situation.

Background information was presented for 12 sec, and the fact was presented for 6 sec, for a total presentation time of 18 sec per story. Stories were presented and then removed from the screen and replaced with a statement containing information pertaining to the fact. Participants evaluated the statement as true or false according to each scenario, using a button press. The question remained on the screen for 4 sec.

To ensure that participants would attend to the final sentence, independent of its moral status, the task on every trial was to answer a ''true/false question'' about the content of the final sentence. We expected participants nevertheless to implicitly evaluate the protagonists while reading the morally relevant information; participants' own reports suggest that they did so. Our previous research confirms that subjects readily make moral judgments of versions of these scenarios (Young & Saxe, 2008; Young et al., 2007). Our hypotheses concerned the time during which participants read either moral or nonmoral facts; therefore, analyses of the neural response were conducted during the ''fact'' time interval.

Participants saw one variation of each scenario, for a total of 48 stories. Stories were presented in a pseudo-random order; the order of conditions was counterbalanced across runs and across participants, ensuring that no condition was immediately repeated. Eight stories were presented in each 4.8-min run; the total experiment, involving six runs, lasted 28.8 min. Fixation blocks of 14 sec were interleaved between each story. The text of the stories was presented in a white 24-point font on a black background. Stories were projected onto a screen via Matlab 5.0 running on an Apple G4 laptop.

In the same scan session, participants participated in four runs of a functional localizer for theory of mind brain regions, contrasting stories that required inferences about a character's beliefs with stories that required inferences about a physical representation, that is, a map or a photograph that has become outdated. Stimuli and story presentation were exactly as described in Saxe and Kanwisher (2003), Experiment 2.

fMRI Analysis

MRI data were analyzed using SPM2 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. Each participant's data were motion corrected and normalized onto a common brain space [Montreal Neurological Institute (MNI) template]. Data were smoothed using a Gaussian filter (FWHM = 5 mm) and were high-pass filtered during analysis. A slow event-related design was used and

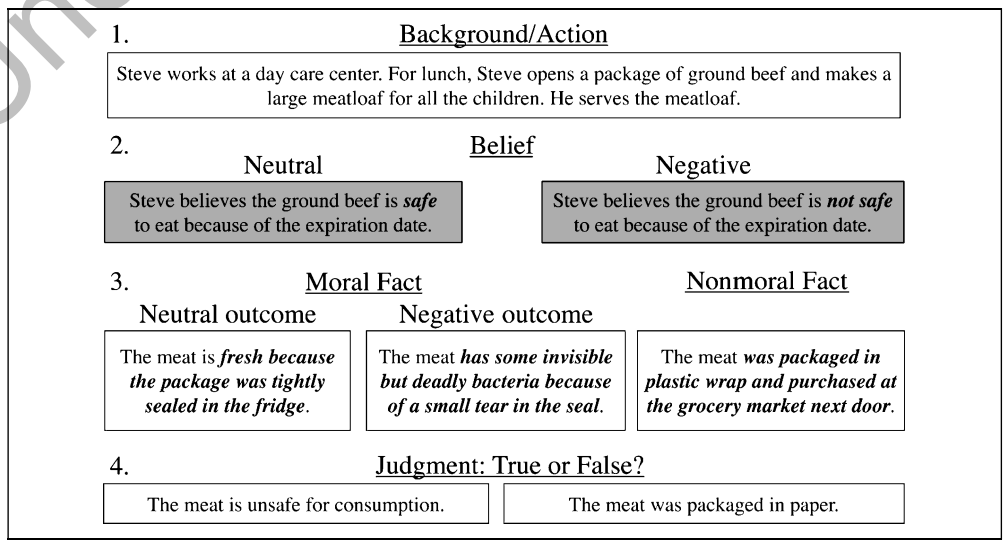

Figure 1. Schematic representation of sample scenario. Changes across conditions are highlighted in bolded and italicized text. In Experiment 1, participants read stories containing the following components. *Background/Action* information was provided to set the scene and contextualize the protagonist's action. *Factual* information revealed either *moral* facts (e.g., whether the protagonist's action would result in a neutral or negative outcome) or *nonmoral* facts about the situation. In Experiment 2, participants read the identical scenarios with the addition of *Belief* information (shaded) following *Background/Action*, which stated whether the protagonist believed her action would result in a neutral or negative outcome. In both experiments, participants made true/false judgments of factual statements. Sentences corresponding to each category were presented in 6-sec blocks.

modeled using a boxcar regressor; an event was defined as a single story, the event onset defined by the onset of text on screen. The timing of the story components was constant for every story so independent parameter estimates could not be created for each component. The response to each component was instead analyzed in the time series extracted from the ROIs (see below).

Both whole-brain and tailored ROI analyses were conducted. Six ROIs were defined for each participant individually based on a whole brain analysis of a localizer contrast and defined as contiguous voxels that were significantly more active ($p < .001$, uncorrected) while the participant read belief as compared with photo stories: RTPJ, LTPJ, PC, dorsal MPFC (dMPFC), middle MPFC (mMPFC), and ventral MPFC (vMPFC). All peak voxels are reported in MNI coordinates (Table 1).

The responses of these ROIs were then measured while participants read moral stories from the current study. Within the ROI, the average percent signal change (PSC) relative to rest baseline (PSC = 100 × BOLD magnitude for (condition − fixation) / average BOLD magnitude during fixation) was calculated for each condition at each time point (averaging across all voxels in the ROI and all blocks of the same condition). PSC during each segment of story presentation (adjusted for hemodynamic lag) in each of the ROIs was compared across experimental conditions. Because the data defining the ROIs were independent from the data used in the repeated measures statistics, Type I errors were drastically reduced.

Results and Discussion

Behavioral Results

RT was analyzed using a repeated measures ANOVA (neutral outcome vs. negative outcome vs. nonmoral), revealing a main effect of fact, $F(2,12) = 22.9$ $p = 8.0 \times 10^{-5}$, partial $h^2 = 0.79$. Post hoc Bonferroni t tests revealed that judgments of nonmoral facts (mean RT = 2.7 sec) took an average of 0.2 sec longer than each of the moral conditions, neutral outcomes (mean RT = 2.5 s), $t(13) = 4.7$, adjusted $p = 1.3 \times 10^{-3}$, and negative outcomes (mean RT = 2.5 s), $t(13) = 6.0$, adjusted $p = 1.4 \times 10^{-4}$. There was no difference between neutral and negative outcomes.

fMRI Results: Localizer Task

To define regions implicated in belief attribution, we contrasted stories that required inferences about a character's beliefs with stories that required inferences about a physical representation such as an outdated photo. A whole-brain random effects analysis of the data replicated results of previous studies using the same task (Saxe & Wexler, 2005; Saxe & Kanwisher, 2003), revealing a higher BOLD response during belief, as compared with photo stories, in the RTPJ, LTPJ, dMPFC, mMPFC, vMPFC, PC, right temporal pole, and right anterior STS ($p < .001$, uncorrected, $k > 10$). ROIs were identified in individual participants (Table 1) at the same threshold: RTPJ (14/14 participants), PC (14/14), LTPJ (14/14), dMPFC (13/14), mMPFC (13/14), and vMPFC (10/14).

fMRI Results: Moral Task

The PSC from rest in each of the ROIs was calculated for the period when the ''fact'' was presented on the screen (16.5–21 sec, accounting for hemodynamic lag). A repeated measures ANOVA (neutral outcome vs. negative outcome vs. nonmoral) revealed a main effect of fact, $F(2,12) = 12.4$ $p = .001$, partial $h^2 = 0.67$, in the RTPJ (Figure 2A and B). In particular, moral facts (mean PSC = 0.26) elicited significantly higher RTPJ activation than nonmoral facts (mean PSC = 0.06), $t(13) = 3.54$ $p = .004$. The RTPJ response was greater for both neu-

Table 1. Localizer Experiment Results

	Experiment 1						*Experiment 2*					
	Individual ROIs			*Whole-Brain Contrast*			*Individual ROIs*			*Whole-Brain Contrast*		
ROI	*x*	*y*	*z*	*x*	*y*	*z*	*x*	*y*	*z*	*x*	*y*	*z*
RTPJ	54	−59	22	54	−52	16	53	−55	21	50	−70	18
PC	0	−61	35	−4	62	32	1	−58	37	6	−62	24
LTPJ	−46	−62	25	−52	−70	26	−54	−58	19	−58	−66	22
dMPFC	2	54	38	0	46	44	3	54	36	0	54	30
mMPFC	2	58	17	2	62	16	4	57	16	0	60	18
vMPFC	2	50	−10	4	50	−4	3	53	−9	4	52	−8

Average peak voxels for ROIs in MNI coordinates for Experiments 1 and 2. The ''Individual ROIs'' columns show the average peak voxels for individual participants' ROIs. The ''Whole-Brain Contrast'' columns show the peak voxel in the same regions in the whole-brain random-effects group analysis.

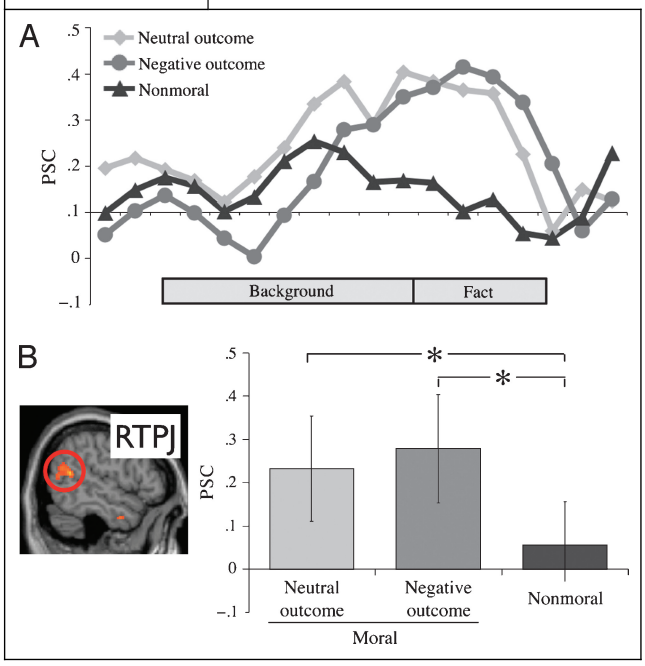

Figure 2. (A) PSC from rest in the RTPJ over time in Experiment 1. Background information was presented for the first 12 sec. Fact was presented for the second 6 sec. Labels have been shifted forward 4.5 sec relative to the stimulus timing to account for hemodynamic lag. (B) PSC at Time 2 (fact) in the RTPJ in Experiment 1. (*Left*) Brain regions where the BOLD signal was higher for (nonmoral) stories about beliefs than (nonmoral) stories about physical representations (n = 14, random effects analysis, $p < .0001$, uncorrected). These data were used to define ROIs, that is, RTPJ. (*Right*) The PSC was significantly greater for moral facts (neutral and negative outcomes, lighter bars) than nonmoral facts (darker bar). Error bars represent standard error.

tral outcomes and negative outcomes than nonmoral facts [neutral outcome vs. nonmoral: $t(13) = 2.16$ $p = .05$; negative outcome vs. nonmoral: $t(13) = 5.10$ $p = 2.0 \times 10^{-4}$]. The RTPJ response did not discriminate between neutral outcomes (mean PSC = 0.23) and negative outcomes (mean PSC = 0.23; $p = .51$). Although we observed an RT difference between moral and nonmoral facts, this RT difference is unlikely to account for the profile of RTPJ recruitment: RTs were not correlated with the magnitude of RTPJ recruitment, across participants, for any condition (for additional evidence, see results for Experiment 2).

Similar patterns were found in the PC and the dMPFC. Paired-samples t tests revealed differences between nonmoral facts and negative outcomes in the PC, $t(13) = 2.16$ $p = .05$, and dMPFC, $t(12) = 2.70$ $p = .02$; however, the responses to neutral outcomes were intermediate and not significantly different from either negative outcomes or nonmoral facts. No significant effects were observed for the LTPJ, mMPFC, or vMPFC.

The results of Experiment 1 suggest that participants may have attempted to spontaneously infer the content of protagonists' mental states (e.g., beliefs). Brain regions associated with the theory of mind, the RTPJ, and to a lesser extent the PC and the dMPFC were recruited selectively for morally relevant outcomes, whether they were neutral or negative. The results support our hypothesis that processing moral stimuli leads to spontaneous inferences about protagonists' beliefs, as no explicit mental state information was provided in the stimuli.

EXPERIMENT 2

In Experiment 2, we reintroduced explicit belief information to further investigate the relationship between theory of mind and moral judgment. If the observed activation in Experiment 1 did indeed reflect spontaneous belief inference, then we might observe a different pattern of response to the very same stimuli (neutral and negative outcomes) when the protagonist's belief was stated explicitly earlier in the stimuli. Previous results are consistent with this latter hypothesis: When explicit belief information was presented first, we observed a higher response to neutral than negative outcomes (Young & Saxe, 2008). In Experiment 2, we therefore added explicit statements of the protagonist's belief to the stories before any information about the outcome. We hypothesized that the response, particularly in the RTPJ, at the time of the outcome would be influenced by whether explicit beliefs were previously presented or not. We hypothesized specifically that the response in theory of mind brain regions at the time of the outcome would be reduced for negative outcomes.

Experiment 2 also provided an opportunity to test two alternative accounts of the main effect of outcome (neutral > negative) previously reported at the time of integration. On the "double-check" hypothesis, because neutral outcomes do not in themselves provide a basis for moral condemnation, the difference reflects *enhanced* belief processing in response to neutral outcomes; subjects therefore double check the belief information in this case (Young & Saxe, 2008). On the "competition" hypothesis, the difference reflects *reduced* belief processing in response to negative outcomes because salient negative outcomes may compete with the processing of previously provided beliefs. Experiment 2 includes a nonmoral fact condition that can serve as a baseline. We can therefore begin to distinguish between these accounts by determining whether the activation for negative outcomes is relatively low ("competition") or whether above-baseline activation is observed for both neutral and negative outcomes, with enhanced activation for neutral outcomes ("double check").

Methods

Fourteen new participants (Harvard College undergraduates, aged 18–22 years, eight women) meeting the same criteria identified in Experiment 1 participated in

a second fMRI experiment. All scan parameters were identical. The stimuli and the task for Experiment 2 were identical as well except in the following regard. Participants read the same scenarios, but explicit statements of beliefs preceded the facts: Protagonists believed their actions would cause harm ("negative" belief) or no harm ("neutral" belief). Stimuli thus consisted of six variations of 48 scenarios for a total of 288 stories. Stories were presented in three cumulative segments, each presented for 6 sec, for a total presentation time of 18 sec per story:

- Background: the protagonist's action in context (identical across conditions).
- Belief: the protagonist's belief about the outcome ("negative" or "neutral").
- Fact: *moral* facts about the action's outcome (negative or neutral) or *nonmoral* facts about the situation.

Each possible belief was true for one outcome and false for the other outcome. All analyses followed the same procedures, as described above for Experiment 1.

Results and Discussion

Behavioral Results

RT was analyzed using a 2 × 3 [Belief (neutral vs. negative) × Fact (neutral vs. negative outcome vs. nonmoral)] repeated measures ANOVA, revealing a main effect of fact, $F(2,12) = 17.5$ $p = 2.8 \times 10^{-4}$, partial $h^2 = 0.75$. Post hoc Bonferroni t tests revealed that RTs for nonmoral facts (mean RT = 2.5 sec) were longer than RTs for neutral outcomes (mean RT = 2.2 s), $t(13) = 6.1$, adjusted $p = 1.1 \times 10^{-4}$. There was no difference between negative outcomes (mean RT = 2.4 sec) and neutral outcomes or nonmoral facts.

fMRI Results: Localizer Task

Exactly as in Experiment 1, regions implicated in belief attribution were functionally localized (Saxe & Kanwisher, 2003). ROIs were identified in individual participants (Table 1): RTPJ (14/14 participants), PC (14/14), LTPJ (12/14), dMPFC (13/14), mMPFC (11/14), and vMPFC (13/14).

fMRI Results: Moral Task

The PSC from rest in the RTPJ was calculated for each of two time intervals:

Time 1 (10.5–15 sec): belief (negative vs. neutral).
Time 2 (16.5–21 sec): fact (neutral outcome vs. negative outcome vs. nonmoral).

During Time 1, belief information was presented. During Time 2, either a morally relevant fact (neutral outcome or negative outcome) was presented or a nonmoral fact was presented. No new belief information was added during Time 2, but prior beliefs (from Time 1) could be integrated with morally relevant outcomes.

At Time 1, no difference in the RTPJ PSC was found between neutral beliefs (mean PSC = 0.23) and negative beliefs (mean PSC = 0.24), $t(13) = 0.41$ $p = .69$. In other words, reading explicit statements of neutral and negative beliefs led to an equally high PSC in the RTPJ. No difference was found in any other ROI for neutral versus negative beliefs.

At Time 2, a 2 × 3 [Belief (neutral vs. negative) × Fact (neutral vs. negative outcome vs. nonmoral)] repeated measures ANOVA revealed a main effect of fact, $F(2,12) = 11.2$ $p = .002$, partial $h^2 = 0.65$, in the RTPJ (Figure 3). Specifically, moral facts (mean PSC = 0.17) elicited higher RTPJ activation than nonmoral facts (mean PSC = 0.02), $t(13) = 3.39$ $p = .005$, as in Experiment 1. However, as predicted and in contrast to the activation pattern observed in Experiment 1, a differential response was found at Time 2: Negative outcomes elicited lower RTPJ activation than neutral outcomes, $t(13) = 4.40$ $p = .001$; 2 × 2 Belief × Outcome repeated measures ANOVA: main effect of outcome, $F(1,13) = 19.38$ $p = .001$, partial $h^2 = 0.60$ (Figure 3).

Similar although less selective patterns were observed in the LTPJ and the PC. As in the RTPJ, a 2 × 2 (Belief × Outcome) repeated measures ANOVA revealed a main effect of outcome in the LTPJ, $F(1,11) = 19.29$ $p = .001$, partial $h^2 = 0.64$, and the PC, $F(1,11) = 7.12$ $p = .02$, partial $h^2 = 0.35$. That is, at Time 2, negative outcomes elicited a lower response than neutral outcomes in both the LTPJ, $t(11) = 4.39$ $p = .001$, and the PC, $t(13) = 2.67$ $p = .02$.

In Experiment 2, when stimuli included explicit belief content, the PSC in theory of mind ROIs, particularly

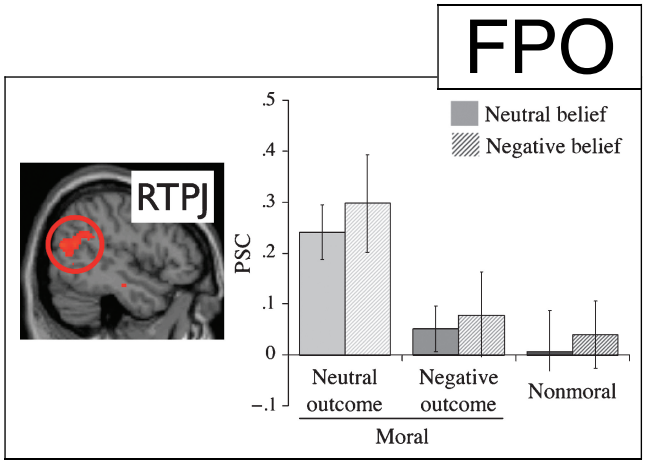

Figure 3. PSC at Time 2 (fact) in the RTPJ in Experiment 2. (*Left*) Brain regions where the BOLD signal was higher for (nonmoral) stories about beliefs than (nonmoral) stories about physical representations ($n = 14$, random effects analysis, $p < .0001$, uncorrected). These data were used to defined ROIs, that is, RTPJ. (*Right*) The PSC was greater for neutral than negative outcomes. Hatched bars represent negative beliefs; filled bars represent neutral beliefs. Error bars represent standard error.

the RTPJ, was greater for moral facts over nonmoral facts. Critically, however, in contrast to the pattern found in Experiment 1, the RTPJ, LTPJ, and PC revealed a further discrimination between neutral outcomes and negative outcomes within moral facts. We note that RT differences were observed only between neutral outcomes and nonmoral facts and thus cannot account for the pattern of neural response, which discriminated between neutral outcomes and negative outcomes. The pattern of results in Experiment 2 (e.g., reduced response for negative as compared with neutral outcomes) exactly replicates our previous findings from a study using a similar design (Young & Saxe, 2008). In both studies, at the time when beliefs *could* be integrated with outcomes, we observed a main effect of neutral over negative outcome in brain regions for belief processing.

Combined Analyses of Experiments 1 and 2

Combined analyses of the PSC at Time 2 (fact) revealed a significant difference in the response profiles for Experiment 1 (belief-absent) and Experiment 2 (belief-present) in the RTPJ. A 2 × 3 [Belief (previously present vs. absent) × Fact (neutral outcome vs. negative outcome vs. nonmoral fact)] mixed ANOVA revealed a significant belief by fact interaction in the RTPJ, $F(2,25) = 8.25$ $p = .002$, partial $h^2 = 0.40$, although in none of the other ROIs. This interaction reflected the fact that the response in the RTPJ to negative outcomes was lower when beliefs were present (Experiment 2) than absent (Experiment 1), $t(26) = 2.66$ $p = .01$, suggestive of a competitive interaction between belief and outcome factors. That is, in the presence of explicit beliefs (Experiment 2), salient *negative* outcome information may compete with the processing of explicit beliefs. Furthermore, in Experiment 2 (belief-present), the level of activation for negative outcomes (mean PSC = 0.06) is comparable to that for nonmoral facts (mean PSC = 0.02; $p = .35$). This pattern suggests reduced belief processing for negative outcomes, as predicted by the ''competition'' hypothesis, rather than enhanced activation for neutral outcomes, as predicted by the ''double-check'' hypothesis.

A random-effects whole brain group analysis ($p > .05$, family-wise correction) of the overall effect of fact (moral vs. nonmoral) and outcome (negative vs. neutral; neutral vs. negative) over the whole block length revealed no significant clusters; this is not surprising, however, because the effects of interest occur at multiple specific times within the trials.

GENERAL DISCUSSION

Moral judgment often represents a response to a complex case characterized by multiple distinct features, such as the action's consequences and the agent's mental states. Mature moral judgment is dominated by mental state information when this information is explicitly available. In real life, however, it is rare for observers to directly access the minds of others. Spontaneous mental state inference, then, presents one cognitive challenge in moral judgment. When mental state information is explicitly available, another cognitive challenge lies in its integration with outcome information. The current study provides neural evidence for both of these processes.

Spontaneous Belief Attribution for Moral Judgment

In the current study, the RTPJ and, to a lesser extent, the PC and the dMPFC regions previously associated with theory of mind (Gobbini et al., 2007; Perner, Aichorn, Kronbichler, Wolfgang, et al., 2006; Saxe & Powell, 2006; Saxe & Wexler, 2005; Saxe & Kanwisher, 2003; Gallagher et al., 2000; Fletcher et al., 1995) were selectively recruited during the presentation of morally relevant facts about actions' outcomes (e.g., effects on other people) but not morally irrelevant facts. This pattern of activation is noteworthy because the task did not explicitly require moral judgments, nor did the stimuli contain explicit mental state information.

These results suggest that under certain conditions, even when no mental state information is given, spontaneous mental state inference occurs. For example, when reading about and perhaps also implicitly evaluating an agent who performs a particular action (e.g., serving meat) that will impact other people in a harmful or harmless way (e.g., the meat is spoiled or fresh), we might attempt to determine the specific content of the agent's beliefs. It may be worth additionally exploring whether neutral outcome information becomes more morally relevant in virtue of its contrast to negative outcome information. For example, reading that ''the meat is fresh'' may provoke consideration of the alternative—the meat *could* have been spoiled. When morally irrelevant facts are described after the action description, no effort is made to establish the mental state of the agent. Morally relevant information, and not just *any* information, therefore appears to afford and perhaps demand spontaneous mental state inference. This finding is broadly consistent with recent research suggesting that morally salient information about an action's outcome prompts backward inferences about the actor's intention (Leslie, Knobe, & Cohen, 2006; Knobe, 2005).

Spontaneous Belief Inference for Nonmoral Judgments

The results of the current study are consistent with other research showing recruitment of theory of mind brain regions for other tasks that do not provide explicit descriptions of beliefs but do require reasoning about actors' unstated beliefs (Sommer et al., 2007). An

example is a nonverbal version of the ''object transfer'' task, which requires participants to determine where an observer will look for an object that was hidden while the observer was either watching or not watching. In this task, greater activation in the RTPJ, in particular, was observed for false belief trials (e.g., the observer looks for the object in the incorrect location, due to a false belief), as compared with true belief trials (e.g., the observer looks in the correct location; Sommer et al., 2007). True belief trials do not necessarily require belief reasoning; participants simply have to respond based on the true location of the object (Dennett, 1978). By contrast, false belief trials in this study require that participants attribute false beliefs to the observer in order both to predict and to explain the observer's behavior (e.g., looking in the wrong place). (Recent behavioral findings suggest that spontaneous belief inference may be even more constrained. Even when an action is performed as a result of an unstated false belief, observers do not always actively represent the belief. Instead, observers maintain a representation of the actor's false belief only as long as the experimental task directly requires them to do so; Apperly, Riggs, Simpson, Samson, & Chiavarino, 2006.) The RTPJ response to false belief trials reported in Sommer et al. (2007) may thus reflect the attribution of beliefs to actors during action prediction and explanation, even in the absence of stimuli containing explicit mental state information (cf. Saxe & Kanwisher, 2003; Experiment 1). Like action prediction, moral judgment also appears to depend on beliefs even if no mental states are stated explicitly. In the current study, we conclude that participants inferred beliefs to explain and to evaluate actions in moral terms.

Integration of Distinct Features for Moral Judgment

In Experiment 2, participants faced a distinct challenge: integrating explicitly stated beliefs with outcomes. Outcome information in Experiment 2 was relevant not only in the sense that it was independently morally relevant, as in Experiment 1, but also in that it rendered the morally relevant belief true or false, thereby potentially directly affecting the representation of the belief.

Although participants read these stimuli, a clear differential response was observed in the RTPJ, LTPJ, and PC: a reduced response for negative as compared with neutral outcomes. This differential response precisely replicates the activation pattern found in a previous study employing a similar design (Young & Saxe, 2008). In both experiments, when explicit information about the belief was presented before information about the outcome, the neural response at the time of the outcome was reduced for negative outcomes. In the current experiment, the response in theory of mind brain regions to negative outcomes was significantly greater when explicit statements of belief were absent than when they were present. These results are consistent with a competitive interaction between belief and outcome factors during moral judgment; salient negative outcomes may compete with the processing of explicit beliefs (regardless of belief valence).

Alternative Explanations of RTPJ Function

Across Experiments 1 and 2, we observed the most robust and selective effects for the RTPJ, although similar patterns were found for the LTPJ, PC, and MPFC, regions for theory of mind. A discussion of other literature suggesting an alternative role for the RTPJ is therefore merited. Lesion and imaging studies implicate the RTPJ in another cognitive task: attentional reorienting in response to unexpected stimuli (Mitchell, 2007; Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000). Nevertheless, the RTPJ response in the current study is best understood as reflecting the representation of mental states for two reasons. First, attentional reorienting cannot explain the highly selective functional response in the RTPJ. In Experiment 1, there was no reason to suspect participants engaged in attentional reorienting on neutral outcomes and negative outcomes (moral facts) but not nonmoral facts. Moral and nonmoral facts were equally frequent and equally expected, but the RTPJ responded selectively while participants were reading moral facts. These conditions were matched for frequency, and, if anything, the pragmatics of the task made morally relevant sentences (both neutral outcomes and negative outcomes) more ''expected'' than the nonmoral and therefore irrelevant facts. This explanation is consistent with postscan accounts from participants: Participants reported reading either morally relevant endings that included the outcome of the protagonist's action or morally irrelevant endings. Furthermore, attentional reorienting cannot explain the selective recruitment of the RTPJ for neutral over negative outcomes in Experiment 2.

Second, a recent study has found that the regions for belief attribution and attentional reorienting are neighboring but distinct (Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, in press). Both individual subject and group analyses revealed less than 8% overlap between the two regions of activation and a reliable separation between the peaks of the two regions: The attention region is located approximately 10 mm superior to the region involved in theory of mind. These results agreed precisely with a recent meta-analysis of 70 published studies that also found that the attention region is 10 mm superior to the region involved in theory of mind (Decety & Lamm, 2007). Given this anatomical separation, the functional localizer approach used in the current study allowed us to identify and then investigate the specific subregion of the RTPJ implicated in theory of mind.

Neuroimaging Data and Reverse Inference

The current study employs ''reverse inference'': The engagement of a cognitive process (belief inference) is inferred from the activation of a particular brain region (RTPJ; Poldrack, 2006; D'Esposito, Ballard, Aguirre, & Zarahn, 1998). Reverse inference, in spite of its known pitfalls, is particularly effective and useful when (1) a brain region has been shown to be selective for a specific cognitive process, and (2) specific behavioral assays have not been established for the cognitive process. Both of these conditions hold in current study.

First, as discussed, of the regions implicated in theory of mind, the RTPJ appears to be the most selective for processing representational mental state content, such as beliefs, both in and outside the moral domain (e.g., Young & Saxe, 2008; Perner, Aichorn, Kronbichler, Wolfgang, et al., 2006; Saxe & Powell, 2006). The RTPJ also revealed the most robust and selective pattern of results in the current study across both experiments. The current study therefore focuses on the RTPJ based on its established selectivity, although it is not the only region involved in theory of mind or moral judgment.

Second, reverse inference is especially useful for determining the presence of a cognitive process for which there are no specific behavioral assays. In the case of spontaneous mental state inference, behavioral assays may lack sensitivity to the underlying cognitive processes. One could, for example, probe subjects for explicit explanations of protagonists' behavior: Do subjects appeal to mental state reasons selectively when explaining morally relevant behavior? However, explicit explanations of any behavior, independent of its moral status, contain mental state reasons, as suggested by previous literature (Malle, 1999) and pilot data on the current stimuli. Experiment 1 therefore benefits from the specific contribution of neuroimaging data and reverse inference.

MPFC and Social Cognition

Although the RTPJ and the PC showed similarly selective patterns across Experiments 1 and 2, the response of the dMPFC showed less a selective pattern, only discriminating between moral and nonmoral facts in Experiment 1 when explicit beliefs were absent. Previous research suggests that the MPFC is recruited not specifically for encoding belief information (e.g., Saxe & Powell, 2006) but more broadly for moral cognition (e.g., Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007; Koenigs et al., 2007; Mendez, Anderson, & Shapira, 2005; Greene et al., 2004) and social cognition (e.g., Mitchell et al., 2006; Adolphs, 2003). Recent work suggests a role for the dMPFC in a specific aspect of theory of mind: judging the desires or valenced attitudes of individuals dissimilar to oneself (Mitchell et al., 2006); by contrast, a more ventral region of MPFC was implicated in judging the desires/attitudes of individuals similar to oneself (Mitchell et al., 2006). It is therefore possible that the dMPFC activation for the stimuli in Experiment 1, which lack explicit beliefs, reflects desire inferences. We emphasize, however, that much more work is needed both to unpack the critical distinction between beliefs versus other mental states, that is, desires, as well as to characterize the functional roles of brain regions that may process different mental state content.

Conclusions

The current study suggests that in the absence of explicit belief information, rather than relying exclusively on outcome information for moral judgment, subjects spontaneously attribute mental states to agents performing morally relevant actions. This result supports the critical role of theory of mind in moral judgment as well as the potential effect that moral judgment may in turn have on spontaneous theory of mind (Cushman, in press; Knobe, 2005). The current study additionally reveals neural signatures of the process by which beliefs are integrated with information about actions' outcomes. These results should thus inform current theories of how beliefs are processed more generally as well as in the specific context of moral judgment.

Acknowledgments

This project was supported by the Athinoula A. Martinos Center for Biomedical Imaging. L. Y. was supported by the NSF. R. S. was supported by MIT and the John Merck Scholars program. Many thanks to Jon Scholz, Susan Carey, Josh Greene, Marc Hauser, and Joshua Knobe for their contributions to this project and Riva Nathans for her help in data collection.

Reprint requests should be sent to Liane Young, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, or via e-mail: lyoung@mit.edu.

REFERENCES

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

Aichorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (in press). Do visual perspective tasks need Theory of Mind. *Journal of Cognitive Neuroscience*.

Apperly, I. A., Riggs, K., Simpson, A., Samson, D., & Chiavarino, C. (2006). Is belief reasoning automatic? *Psychological Science, 17,* 841–844.

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development, 103,* 37–49.

Baird, J. A., & Moses, L. J. (2001). Do preschoolers appreciate that identical actions may be motivated by different intentions? *Journal of Cognition and Development, 2,* 413–448.

Borg, J. S., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience, 18,* 803–817.

Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 2,* 84–92.

Ciaramidaro, A., Adenzato, M., Enrici, I., Erk, S., Pia, L., Bara, B. G., et al. (2007). The intentional network: How the brain reads varieties of intentions. *Neuropsychologia, 45,* 3105–3113.

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience, 3,* 292–297.

Cushman, F. (in press). Crime and punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition.*

Cushman, F., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuitions in moral judgment: Testing three principles of harm. *Psychological Science, 17,* 1082–1089.

Dennett, D. (1978). Beliefs about beliefs. *Behavioral and Brain Science, 1,* 568–570.

D'Esposito, M., Ballard, D., Aguirre, G., & Zarahn, E. (1998). Human prefrontal cortex is not specific working memory: A functional MRI study. *Neuroimage, 8,* 274–282.

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

Gallagher, H. L., Happe, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of ''theory of mind'' in verbal and nonverbal tasks. *Neuropsychologia, 38,* 11–21.

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience, 19,* 1803–1814.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44,* 389–400.

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.

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences, 9,* 357–359.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446,* 908–911.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: ''Theory of mind'' and moral judgment. *Psychological Science, 6,* 421–427.

Malle, B. (1999). How people explain behavior: A new theoretical framework. *Personality and Psychology Review, 3,* 23–49.

Mendez, M., Anderson, E., & Shapira, J. (2005). An investigation of moral judgment in frontotemporal dementia. *Cognitive and Behavioral Neurology, 18,* 193–197.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11,* 143–152.

Mitchell, J. P. (2007). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*.

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

Perner, J., Aichorn, M., Kronbichler, M., Wolfgang, S., & Laddurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience, 1,* 235–2258.

Perner, J., Aichorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience, 1,* 245–258.

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10,* 59–63.

Ruby, P., & Decety, J. (2003). What you belief versus what you think they believe: A neuroimaging study of conceptual perspective-taking. *European Journal of Neuroscience, 17,* 2475–2480.

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.

Saxe, R., & Powell, L. (2006). It's the thought that counts: Specific brain regions for one component of Theory of Mind. *Psychological Science, 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.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E., & Saxe, R. (in press). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention.

Sommer, M., Dohnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. *Neuroimage, 35,* 1378–1384.

Vogeley, K., Bussfield, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., et al. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *Neuroimage, 14,* 170–181.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, U.S.A., 104,* 8235–8240.

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *Neuroimage, 40,* 1912–1920.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development, 67,* 2478–2492.