ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 40 (2008) 1912–1920

The neural basis of belief encoding and integration in moral judgment

Liane Young[a,b,*] and Rebecca Saxe[b]

[a]Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA
[b]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, USA

Received 6 November 2007; revised 13 January 2008; accepted 19 January 2008
Available online 14 February 2008

Moral judgment in the mature state depends on "theory of mind", or the capacity to attribute mental states (e.g., beliefs, desires, and intentions) to moral agents. The current study uses functional magnetic resonance imaging (fMRI) to investigate the cognitive processes for belief attribution in moral judgment. Participants read vignettes in a 2×2×2 design: protagonists produced either a negative or neutral outcome, based on the belief that they were causing the negative outcome or the neutral outcome; presentation of belief information either preceded or followed outcome information. In each case, participants judged the moral permissibility of the action. The results indicate that while the medial prefrontal cortex is recruited for processing belief valence, the temporo-parietal junction and precuneus are recruited for processing beliefs in moral judgment via two distinct component processes: (1) encoding beliefs and (2) integrating beliefs with other relevant features of the action (e.g., the outcome) for moral judgment.

Keywords: Morality; Theory of mind; Belief attribution; fMRI; Temporo-parietal junction; Precuneus; Medial prefrontal cortex

Introduction

One key cognitive input to moral judgment is "theory of mind" or the capacity to attribute mental states, such as beliefs, desires, and intentions, to moral agents (e.g., Baird and Astington, 2004; Borg et al., 2006; Cushman et al., 2006; Knobe, 2005; Mikhail, 2007; Young et al., 2007). Adults judge intentional harms to be morally worse than the same harms brought about accidentally or unknowingly. In the current study, we investigate the neural evidence for multiple distinct cognitive processes underlying theory of mind in moral judgment.

The neural basis of theory of mind has been investigated in recent functional magnetic resonance imaging (fMRI) studies. These studies reveal a consistent group of brain regions for "theory of mind" in nonmoral contexts: the medial prefrontal cortex, right and left temporo-parietal junction, and precuneus (Ciaramidaro et al., 2007; Fletcher et al., 1995; Gallagher et al., 2000; Gobbini et al., 2007; Ruby and Decety, 2003; Saxe and Kanwisher, 2003; Vogeley et al., 2001). Of these regions, the right temporo-parietal junction (RTPJ) in particular appears to be selective for belief attribution (Aichorn et al., 2006; Fletcher et al., 1995; Gallagher et al., 2000; Gobbini et al., 2007; Perner et al., 2006; Saxe and Wexler, 2005; Sommer et al., 2007). For example, its response is high when subjects read stories that describe a character's thoughts and beliefs but low during stories containing other socially relevant information (e.g., a character's physical appearance, cultural background, or even internal subjective sensations such as hunger or fatigue; Saxe and Powell, 2006).

A recent fMRI study showed that these same brain regions are recruited for moral judgment, particularly, judgment of intentional and unintentional harms and non-harms (Young et al., 2007). These brain regions showed significant activation above baseline for all conditions of moral judgment but were modulated by an interaction between mental state and outcome factors. In the current study, we sought to refine our characterization of the role of these brain regions. Evidence from developmental psychology suggests that the acquisition of the theory of mind skills required for mature moral judgment is marked by multiple distinct cognitive achievements. We investigated whether these different developmental stages correspond to distinct functional profiles in the adult brain.

The classic task for assessing a child's ability to reason about the mental states of others is the false belief task (for a review, see Flavell, 1999; Wellman et al., 2001). In its standard version, known as the "object transfer" problem, the child is told a story in which a character's belief about the location of a target object becomes false when the object is moved without the character's knowledge. Generating the correct answer requires the child to pay attention to the character's belief, not just to the true location of the object. While the precise age of success varies between children and between versions of the task, in general, children younger than 3 or

* Corresponding author. Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA. Fax: +1 617 258 8654.
E-mail address: lyoung@fas.harvard.edu (L. Young).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.01.057

4 years old cannot verbalize correct answers to false belief problems (but see Onishi and Baillargeon, 2005). By the time they are five, children reliably pass the false belief test.

This capacity appears to precede rather than to coincide with the capacity to use belief information in the context of moral judgment. Five year olds can make moral distinctions based on mental state distinctions only when consequences are held constant (Karniol, 1978; Nelson Le Gall, 1985; Nunez and Harris, 1998; Siegel and Peterson, 1998; Wellman et al., 1979). Even though they can represent beliefs, these children continue to base their moral judgments primarily on the action's consequences rather than the actor's beliefs, when these two factors conflict (Hebble, 1971; Piaget, 1965/1932; Shultz et al., 1986; Yuill, 1984; Yuill and Perner, 1988; Zelazo et al., 1996). For example, five year olds judge that an agent who intends to direct a traveler to the right location but accidentally misdirects him is worse than another agent who intends to misdirect a traveler but accidentally directs him to the right place (Piaget, 1965/1932). Only later are children able to generate adult-like judgments of these scenarios, which continue to take consequences into account (Cushman, personal communication) but additionally depend substantially on beliefs (Baird and Astington, 2004; Baird and Moses, 2001; Darley and Zanna, 1982; Fincham and Jaspers, 1979; Karniol, 1978; Shultz et al., 1986; Yuill, 1984) thereby requiring true integration of information about consequences and beliefs (Grueneich, 1982; Weiner, 1995; Zelazo et al., 1996).

Based on this evidence from developmental psychology, we propose a distinction between two separate component processes of belief attribution in moral judgment: encoding and integration. Encoding consists of forming an initial representation of the protagonist's belief. Integration, by contrast, consists of using the belief for moral judgment in flexible combination with relevant outcome information. On this analysis, five-year-old children are capable of encoding beliefs (e.g., in the false belief task), but they cannot fully integrate beliefs with outcomes in the service of moral judgment. Here we investigate the neural evidence for these cognitive processes in the adult brain. We suggest that the brain regions for encoding should be (1) recruited when belief information is first presented and (2) recruited selectively for belief information over non-belief information. As such, the response at encoding should be stimulus-bound, that is, modulated by whether the current stimulus being processed contains belief content. Brain regions for integration should be (1) recruited once morally relevant non-belief information (e.g., outcome) is available and (2) show a functional profile reflecting the interaction between belief and outcome. The response at integration should therefore reflect the use of prior belief information in constructing a moral judgment and the influence of outcome information on belief processing.

In the current study, participants read vignettes in a $2\times2\times2$ design (Fig. 1): protagonists produced either a negative outcome or a neutral outcome, based on the belief that they were causing the negative outcome ("negative" belief) or the neutral outcome ("neutral" belief); belief information could be presented either before or after information foreshadowing the outcome. A protagonist with a negative belief who produced a negative outcome did so knowingly, while a protagonist with a negative belief who produced a neutral outcome did so unknowingly or accidentally, based on a false belief. In each case, participants judged the moral permissibility of the protagonist's action. This design allowed us to address the following questions with respect to theory of mind in mature moral judgment: (1) Is there neural evidence for encoding and integration as distinct processes? (2) Are brain regions previously implicated in belief attribution in nonmoral contexts specifically involved in belief encoding and/or belief integration? (3) If so, are encoding and integration accomplished by the same or different subsets of these regions?

Materials and methods

Seventeen naive right-handed subjects (Harvard College undergraduates, aged 18–22 years, six women) participated in the functional MRI study for payment. All subjects were native English speakers, had normal or corrected-to-normal vision and gave written informed consent in accordance with the requirements of the internal review board at MIT. Subjects were scanned at 3-T (at the MIT scanning facility in Cambridge, Massachusetts) using twenty-six 4-mm-thick near-axial slices covering the whole brain. Standard echoplanar imaging procedures were used (TR = 2 s, TE = 40 ms, flip angle 90°).

Stimuli consisted of eight variations of 48 scenarios for a total of 384 stories with an average of 86 words per story (see Supplementary data for full text of scenarios). A $2\times2\times2$ design was used for each scenario: (i) protagonists produced either a negative outcome (harm to a person) or a neutral outcome (no harm); (ii) protagonists held the belief that they were causing a negative outcome ("negative" belief) or a neutral outcome ("neutral" belief); (iii) either belief information or information foreshadowing the outcome was presented first. Stories were presented in four cumulative segments, each presented for 6 s, for a total presentation time of 24 s per story:

(1) Background: information to set the scene (identical across all conditions)
(2 or 3) Foreshadow: information foreshadowing the outcome (negative or neutral)
(2 or 3) Belief: the protagonist's belief about the situation (negative or neutral)
(4) Outcome: the protagonist's action and its outcome (negative or neutral)

For example, as in the scenario in Fig. 1, the identification of the white powder by the coffee as poison rather than sugar foreshadows a person's death by poison. In every story used in this experiment, when something is wrong at this stage (e.g., poison in place of sugar, drowning swimmer, asthma attack), the protagonist's action or inaction results in a negative outcome (someone's death). Each possible belief was true for one outcome and false for the other outcome. Stories were presented and then removed from the screen and replaced with a question about the moral nature of the action. Participants were asked to make judgments on a scale of 1 (forbidden) to 3 (permissible), using a button press. Three buttons were used due to the malfunction of a fourth button on the scanner-safe response apparatus. The question remained on the screen for 4 s.

Subjects saw one variation of each scenario, for a total of 48 stories. Stories were presented in a pseudorandom order, the order of conditions counterbalanced across runs and across subjects, thereby ensuring that no condition was immediately repeated. Eight stories were presented in each 5.6 min run; the total experiment, involving six runs, lasted 33.6 min. Fixation blocks of 14 s 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.

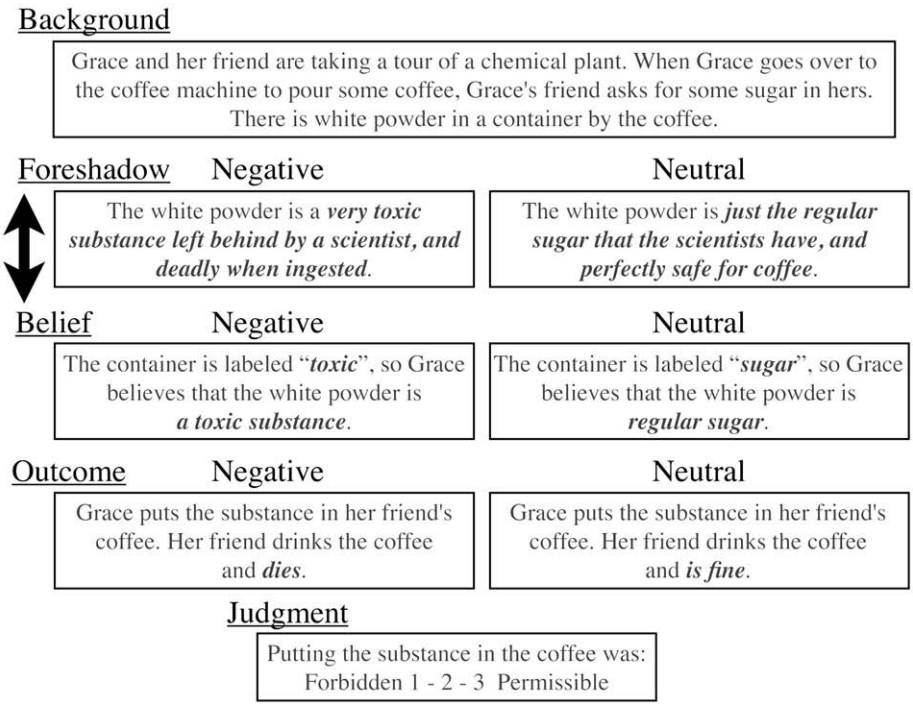

Fig. 1. Schematic representation of sample scenario. Changes between conditions are highlighted in bold italics. "Foreshadow" information foreshadows whether the action will result in a negative or neutral outcome. "Belief" information states whether the protagonist holds a belief that he or she is in a negative situation and that action will result in a negative outcome (negative belief) or a belief that he or she is in a neutral situation and that action will result in a neutral outcome (neutral belief). During belief-first trials, belief information was presented first and foreshadow information was presented second. During foreshadow-first trials, foreshadow information was presented first and belief information was presented second. Sentences corresponding to each category were presented in 6 s blocks.

In the same scan session, subjects participated in four runs of a localizer experiment, contrasting stories that required inferences about a character's beliefs (belief condition) with stories that required inferences about a physical representation, i.e. a photo that has become outdated (photo condition). Stimuli and story presentation were exactly as described in Saxe and Kanwisher (2003), Experiment 2.

fMRI data analysis

MRI data were analyzed using SPM2 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. Each subject's data were motion corrected and then normalized onto a common brain space (the Montreal Neurological Institute, MNI, template). Data were then smoothed using a Gaussian filter (full width half maximum = 5 mm), and high-pass filtered during analysis. A slow event-related design was used and modeled using a boxcar regressor. An event was defined as a single story (30 s); the event onset was defined by the onset of text on the screen. The timing of the four story components was constant for every story; thus, independent parameter estimates were not created for each component. Components were separated by the time of response, accounting for the hemodynamic lag.

Both whole-brain and tailored regions of interest (ROI) analyses were conducted. Six ROIs were defined for each subject individually based on a whole-brain analysis of a localizer contrast, and defined as contiguous voxels that were significantly more active ($p < 0.001$, uncorrected) while the subject read belief stories, as compared with photo stories: RTPJ, LTPJ, PC, dMPFC, mMPFC, and vMPFC. All peak voxels are reported in Montreal Neurological Institute coordinates.

The responses of these regions of interest were then measured while subjects read stories from the current experiment. Within the ROI, the average percent signal change (PSC) relative to rest baseline (PSC = 100 × raw BOLD magnitude for (condition − fixation) / raw BOLD magnitude for 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 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

Subjects evaluated the moral status of protagonists' actions using three buttons associated with a scale from completely forbidden (1) to completely permissible (3). To determine the effects of belief and outcome and order, we used a 2 × 2 × 2 (outcome [negative vs. neutral] by belief ["negative" vs. "neutral"] by order [belief-first vs. foreshadow-first]) repeated measures ANOVA. Actions performed by protagonists with "negative" beliefs were judged to be less permissible than actions performed by protagonists with "neutral" beliefs (negative: 1.2, neutral: 2.2; $F(1,11) = 69.7$, $p = 4.4 \times 10^{-6}$, partial $\eta^2 = 0.86$). Actions resulting in negative outcomes were judged to be less permissible than actions resulting in neutral outcomes (negative: 2.1, neutral: 2.5; $F(1,11) = 20.4$, $p = 0.001$, partial $\eta^2 = 0.65$). No other main effect or interaction achieved significance. The same

2×2×2 repeated measures ANOVA was performed for reaction time, yielding no significant main effects or interactions.

fMRI results: localizer task

To define regions implicated in belief attribution, stories that required inferences about a character's beliefs (belief condition) were contrasted with stories that required inferences about a physical representation such as an outdated photograph (photo condition). A whole-brain random effects analysis of the data replicated results of previous studies using the same task (Saxe and Kanwisher, 2003; Saxe and Wexler, 2005), revealing a higher BOLD response during belief, as compared to photo stories, in the RTPJ, LTPJ, dorsal (d), middle (m), and ventral (v) MPFC, precuneus (PC), right temporal pole, and right anterior superior temporal sulcus ($p<0.001$, uncorrected, $k>10$). Regions of interest (ROIs) were identified in individual subjects (Table 1) at the same threshold: RTPJ (15/17 subjects), PC (17/17), LTPJ (16/17), dMPFC (14/17), mMPFC (12/17), and vMPFC (10/17).

fMRI results: moral judgment task

The average percent signal change (PSC) from rest in each region of interest was calculated for each of three time intervals:

Time 1 (10–14 s): belief (belief-first trials) or foreshadow (foreshadow-first trials)
Time 2 (16–20 s): foreshadow (belief-first trials) or belief (foreshadow-first trials)
Time 3 (22–26 s): information about the protagonist's action

Times 1 and 2 represent the time during which the *encoding* of the belief may occur; belief information is being presented for the first time, and information relevant for moral judgment is incomplete. Time 3 represents the time during which the *integration* of the belief may occur; no new belief information is added, but prior belief information may be integrated with information about the protagonist's action and the actual outcome in the construction of a moral judgment.

Encoding

The PSCs for the earlier times, times 1 and 2 (encoding), during which belief and foreshadow information were initially presented, were analyzed using a 2×2×2×2 (time [1 vs. 2] by outcome [negative vs. neutral] by belief ["negative" vs. "neutral"] by order [belief-first vs. foreshadow-first]) repeated measures ANOVA.

Table 1
Localizer experiment results

ROI	Individual ROIs			Whole-brain contrast		
	x	*y*	*z*	*x*	*y*	*z*
RTPJ	56	−56	22	56	−54	28
PC	−1	−58	39	−2	−60	40
LTPJ	−50	−63	26	−52	−58	26
dMPFC	−2	58	29	2	60	28
mMPFC	1	59	15	−4	56	8
vMPFC	1	55	−7	0	54	−8

Average peak voxels for ROIs in Montreal Neurological Institute coordinates. The "Individual ROIs" columns show the average peak voxels for individual subjects' ROIs. The "Whole-brain contrast" columns show the peak voxel in the same regions in the whole-brain random effects group analysis.

(1) RTPJ: A significant time by order interaction was observed in the RTPJ ($F(1,14)=8.0$, $p=0.01$): the average response was higher at time 1 when belief information (mean PSC: 0.41) was presented at time 1, than when foreshadow information (mean PSC: 0.35) was presented at time 1; and higher at time 2 when belief (mean PSC: 0.54) was presented at time 2 than when foreshadow (mean PSC: 0.44) was presented at time 2. Planned comparisons at times 1 and 2 did not yield significant differences between belief and foreshadow, though averaging over times 1 and 2 revealed a greater response for belief than foreshadow (mean belief PSC: 0.47; mean foreshadow PSC: 0.40; $t(14)=2.82$, $p=0.01$). The PSC in the RTPJ therefore appeared to track with whether the stimulus being presented contained belief information or not (Fig. 2, top panel; Table 2). However, the response during the encoding of the belief did not depend on the content or "valence" of the belief. At the time that the belief was presented, there was no difference between the responses to "negative" versus "neutral" belief (belief at time 1: "negative": 0.39, "neutral": 0.41, $t(14)=0.32$, $p=0.76$; belief at time 2: "negative": 0.55, "neutral": 0.53, $t(14)=-0.32$, $p=0.75$). There were also no main effects of negative versus neutral foreshadow during encoding.
(2) PC and LTPJ: The PC and LTPJ showed a similar though less selective profile at encoding (Tables 2 and 3, Supplementary Fig. 1). A time (1 vs. 2) by order (belief-first vs. foreshadow-first) interaction was observed in both the PC ($F(1,16)=7.4$, $p=0.02$) and the LTPJ ($F(1,15)=5.0$, $p=0.04$). That is, the response in the PC was higher at time 1 when belief information was presented at time 1 and higher at time 2 when belief information was presented at time 2, suggesting that the response in the PC during encoding, like the RTPJ, is driven by the stimulus—whether the stimulus contains belief information. Like the RTPJ, planned comparisons for the PC at times 1 and 2 did not yield significant differences between belief and foreshadow, though averaging over times 1 and 2 revealed a greater response for belief than foreshadow (mean belief PSC: 0.07; mean foreshadow PSC: 0.001; $t(16)=2.71$, $p=0.02$). The interaction was less selective in the LTPJ: belief versus foreshadow was discriminated at time 2 but not at time 1.
(3) MPFC: Regions in the MPFC showed a different pattern from the RTPJ, PC, and LTPJ. There was no evidence that any region of the MPFC was recruited for belief encoding (Fig. 2, bottom panel). No significant main effects or interactions were found during times 1 and 2 in the dMPFC, mMPFC, or vMPFC. To determine whether the profile found during encoding (times 1 and 2) for the RTPJ (e.g., time by order interaction) was significantly different from the profile found for regions in the MPFC, a 2×2×2 repeated measures ANOVA was conducted for every pair of regions that included the RTPJ and one region in the MPFC. The predicted time by order by region interactions were significant ($p<0.05$) in all pairs.

Integration

The PSC for time 3 (integration) was analyzed using a 2×2×2 (outcome [negative vs. neutral] by belief ["negative" vs. "neutral"] by order [belief-first vs. foreshadow-first]) repeated measures

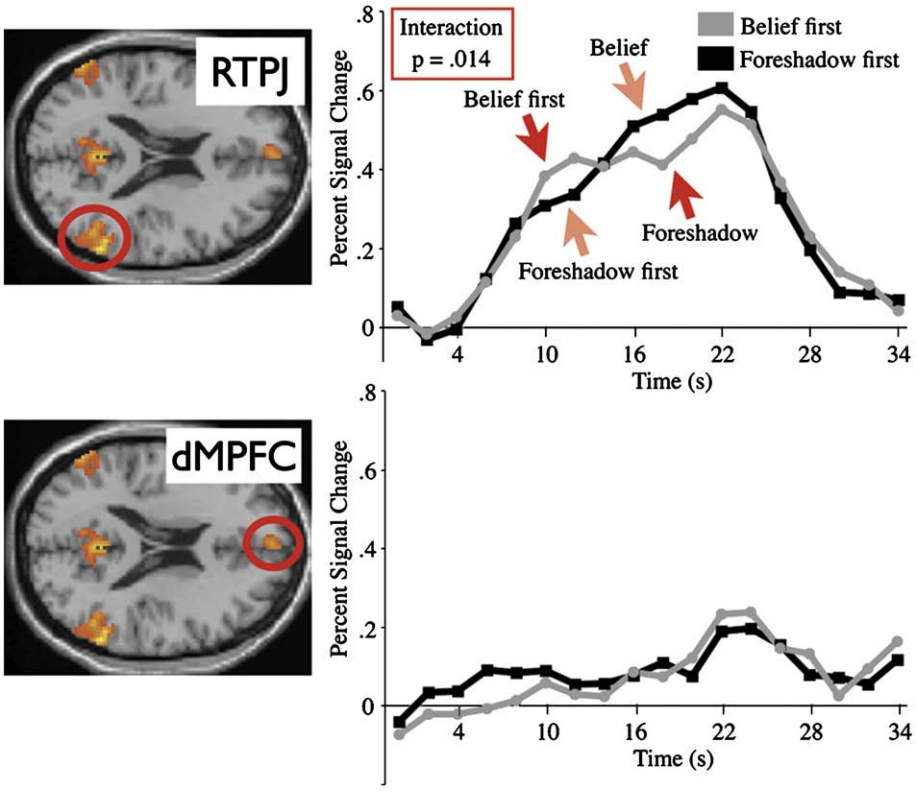

Fig. 2. PSC from rest in the RTPJ (*top*) and dMPFC (*bottom*) over time. (*Left*) Brain regions where the BOLD signal was higher for (nonmoral) stories about beliefs than (nonmoral) stories about physical representations ($N=17$, random effects analysis, $p<0.001$ uncorrected). These data were used to define ROIs: RTPJ (*top*), dMPFC (*bottom*). (*Right*) The PSC in the RTPJ (*top*) and dMPFC (*bottom*) during belief-first trials (gray) and foreshadow-first trials (black). Time 1 (*10–14 s*): belief information was presented during belief-first trials; foreshadow information was presented during foreshadow-first trials. Time 2 (*16–20 s*): foreshadow information was presented during belief-first trials; belief information was presented during foreshadow-first trials. Time 3 (*22–26 s*): information was presented about the protagonist's action and the outcome.

ANOVA (Table 3). At time 3, the protagonist's action, the subject of moral judgment, and its actual outcome were described.

(1) RTPJ: Even though no new belief information was presented, the PSC in the RTPJ was significantly above baseline in all eight conditions ($p<0.01$). Also, a significant outcome by belief by order interaction ($F(1,14)=17.0$, $p=0.001$, partial $\eta^2=0.55$) was found, suggesting that the contribution of the factors of belief and outcome depended on the order of information presentation (belief-first vs. foreshadow-first). Each order was therefore analyzed separately.

For foreshadow-first, the response showed a main effect of "negative" belief over "neutral" belief ($F(1,14)=9.7$, $p=0.008$) and a belief by outcome interaction ($F(1,14)=11.2$, $p=0.005$; Fig. 3, top panel, as reported in Young et al., 2007, Experiment 2). Planned comparisons revealed that the PSC was higher for "negative" belief than "neutral" belief in the case of a neutral outcome ("negative": 0.74, "neutral": 0.32, $t(14)=4.0$, $p=0.001$), but was not significantly different for "negative" and "neutral" belief in the case of a negative outcome ("negative": 0.41, "neutral" PSC: 0.51, $t(14)=-1.3$ $p=0.22$). Post-hoc Bonferroni's *t*-tests revealed that the PSC for attempted harm was significantly greater than each of the other conditions (unknowing harm: $t(14)=2.6$, adjusted $p=0.04$; intentional harm: $t(14)=-3.0$, adjusted $p=0.02$). Consistent with the regions of interest analysis, a random effects whole-brain analysis ($p>0.001$, uncorrected) revealed greater activation for attempted harm (negative belief, neutral outcome) as compared to all-neutral stories in the RTPJ, for this order (average peak voxel coordinates [48 −46 16]).

Table 2
Mean PSC in three ROIs during times 1 and 2 of the moral scenarios

ROI	Mean PSC				Interaction of time × order			
	Belief-first		Foreshadow-first					
	Time 1	Time 2	Time 1	Time 2	*df*	*F*	*p* value	Partial η^2
RTPJ	0.41	0.44	0.35	0.54	(1,14)	8.00	0.01	0.36
PC	−0.02	0.06	−0.058	0.16	(1,16)	7.40	0.02	0.32
LTPJ	0.33	0.39	0.35	0.51	(1,15)	5.00	0.04	0.25

All three of the regions showed a significant interaction between time (time 1 vs. time 2) and order (belief-first vs. foreshadow-first).

Table 3
Mean PSC in three ROIs during time 3 of the moral scenarios

ROI	Mean PSC (belief, outcome)				Interaction of belief × outcome			
	Neut, Neut	Neut, Neg	Neg, Neut	Neg, Neg	*df*	*F*	*p* value	Partial η^2
RTPJ	0.32	0.51	0.74	0.41	(1,14)	11.20	0.01	0.44
PC	0.07	0.18	0.29	0.11	(1,16)	7.20	0.02	0.31
LTPJ	0.22	0.4	0.56	0.39	(1,15)	5.00	0.04	0.25

All three of the regions showed a significant interaction between negative (Neg) and neutral (Neut) belief and outcome information.

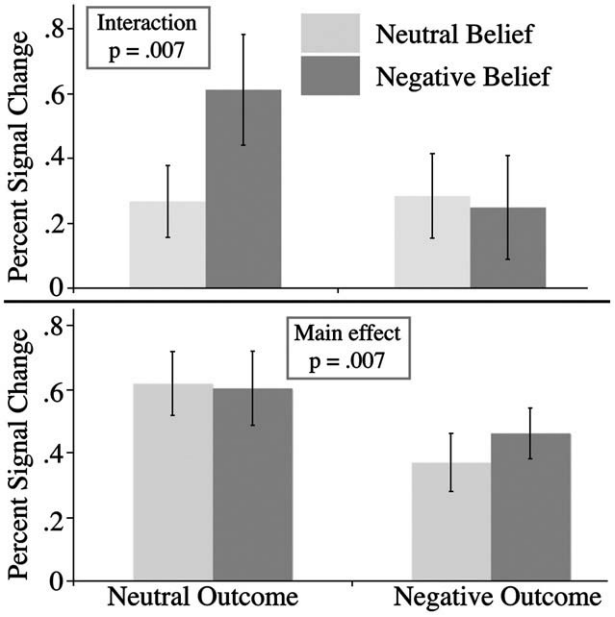

Fig. 3. PSC from rest in the RTPJ during time 3 (22–26 s). Error bars correspond to standard error. (*Top*) Foreshadow-first trials. (*Bottom*) Belief-first trials.

A different pattern was observed for belief-first trials (Fig. 3, bottom panel). There was no interaction between belief and outcome. Only a main effect of neutral outcome over negative outcome was significant (neutral: 0.60, negative: 0.36, $F(1,14)=10.2$, $p=0.007$, partial $\eta^2=0.42$). Consistent with this effect, a whole-brain analysis ($p>0.001$, uncorrected) of the overall effect of outcome (neutral over negative) revealed activation in the RTPJ (average peak voxel coordinates [60 −58 20]).

(2) PC and LTPJ: The same 2×2×2 (outcome [negative vs. neutral] by belief ["negative" vs. "neutral"] by order [belief-first vs. foreshadow-first]) repeated measures ANOVA was conducted for the PC and LTPJ at time 3. A significant belief by outcome by order interaction was found only in the PC ($F(1,16)=5.7$, $p=0.03$). Separate analyses for both orders were performed for both the PC and LTPJ.

A belief by outcome interaction for foreshadow-first was found in the PC ($F(1,16)=7.2$, $p=0.02$). Planned comparisons revealed that the PSC in the PC was higher for "negative" belief than "neutral" belief in the case of a neutral outcome ("negative": 0.29, "neutral": 0.07, $t(16)=3.24$, $p=0.005$), but was not significantly different for "negative" and "neutral" belief in the case of a negative outcome ("negative": 0.11, "neutral" PSC: 0.17, $t(16)=-0.73$, $p=0.48$).

The LTPJ showed the same belief by outcome interaction for foreshadow-first ($F(1,15)=5.0$, $p=0.04$) as well as a main effect of "negative" over "neutral" belief ($F(1,15)=8.5$, $p=0.01$). Planned comparisons revealed that the PSC in the LTPJ was higher for "negative" belief than "neutral" belief in the case of a neutral outcome ("negative": 0.56, "neutral": 0.22, $t(15)=3.73$, $p=0.002$), but was not significantly different for "negative" and "neutral" belief in the case of a negative outcome ("negative": 0.39, "neutral" PSC: 0.40, $t(15)=-0.10$, $p=0.92$).

In contrast to the RTPJ, neither the PC nor the LTPJ showed a main effect of neutral outcome over negative outcome (or any other significant main effects or interactions) for belief-first trials.

(3) MPFC: The belief by outcome by order interaction was not significant in any region of the MPFC. No main effects or interactions were found for the mMPFC or the vMPFC. The dMPFC, however, showed a main effect of "negative" over "neutral" belief ($F(1,13)=9.4$, $p=0.01$) for foreshadow-first trials, suggesting a unique role for the dMPFC in processing belief valence for moral judgment. In a previous study using similar stimuli (Young et al., 2007), we observed a similar trend in the dMPFC that did not reach significance (negative>neutral belief, $p<0.1$). However, those results were based on an analysis of only nine individuals. To further investigate the reliability of this effect, we therefore analyzed the response in the dMPFC at the time of integration, across both experiments. A 2×2 (belief by outcome) ANOVA ($N=23$) revealed a strong main effect of belief (negative>neutral belief, $F(1,22)=13.3$, $p=0.001$, partial $\eta^2=0.38$), although there was also a significant interaction between belief and outcome ($F(1,22)=11.4$, $p=0.003$, partial $\eta^2=0.34$), similar to that observed in the other regions investigated.

General discussion

Moral judgment in the mature state depends on the capacity to attribute beliefs to agents. Both previous and current results suggest that, when belief and outcome information conflict, adult moral judgments are determined primarily by the belief (Cushman, personal communication; Young et al., 2007). Here we distinguish between two cognitive processes associated with belief attribution in moral judgment: the encoding and integration of beliefs. First, belief information is encoded; that is, the relevant belief is detected and represented. Second, belief information is integrated with other relevant information in the construction of moral judgment; belief information is represented in terms of its relation to outcome information. Our results suggest that the same brain regions, the RTPJ, PC, and LTPJ, support both of these belief processes, reflecting a differential response during both encoding and integration phases. The dMPFC, by contrast, appears to process belief valence for moral judgment during the integration phase. Thus, while the RTPJ, PC, and LTPJ are responsible for processing beliefs for moral judgment, the dMPFC is responsible for processing an explicitly morally relevant feature of the action: whether the actor believed he or she was causing harm. Here, we investigate the functional profiles of the response in these regions during moral judgment.

The current study reveals neural signatures of the process by which belief information is encoded. This process appears to be supported by the RTPJ, the PC and, to a lesser extent, the LTPJ. Recruitment of these brain regions was observed early in the stimulus, when subjects were first presented with information about the protagonist's belief. This response was selective for explicit belief information in the current stimulus, as revealed by a significant time by order (belief-first vs. foreshadow-first trials) interaction (cf. Saxe and Wexler, 2005) and consistent with previous research supporting the specific role of these regions but, in particular, the RTPJ, in processing beliefs (Aichorn et al., 2006; Fletcher et al., 1995; Gallagher et al., 2000; Gobbini et al., 2007; Perner et al., 2006; Saxe and Kanwisher, 2003; Saxe and Powell, 2006; Saxe and Wexler, 2005). Interestingly, the response in these regions at encoding was not influenced by the valence or content of the belief (i.e. "negative" vs. "neutral") in any of the regions tested.

During the integration phase, when subjects were able to make moral judgments of the protagonist's action and its outcome, the response in the RTPJ, the PC, and LTPJ showed a different functional profile. During this time, no new belief information was added to the story; however, these regions showed above-baseline recruitment

that differentiated among different moral conditions based on aspects of both the belief and the outcome. The response after the presentation of belief information may reflect the integration of previously presented belief information with other task-relevant information in constructing a coherent moral judgment (Grueneich, 1982; Weiner, 1995; Zelazo et al., 1996). In the context of the current study, outcome information is relevant in two senses: 1) outcome information renders the morally relevant belief true or false, thereby affecting the representation of the belief, and 2) outcome information is independently morally relevant insofar as we judge harms worse than non-harms, and therefore must be reconciled with morally relevant belief information. The current study conflates these two senses of relevance by using outcome information in a moral context, but this distinction should be explored in future studies. Furthermore, while we have focused on the encoding–integration distinction in the context of moral judgment of actions that result in harms or non-harms, it would be of interest to determine what other morally relevant information demands integration with belief information, and whether the same encoding–integration distinction appears in nonmoral domains.

The specific functional profile observed during integration differed across brain regions and across stimulus orders. Replicating previous research (Young et al., 2007), we observed a belief by outcome interaction in the RTPJ, PC, and LTPJ when foreshadow information had been presented before belief information. Furthermore, the RTPJ response is significantly higher in the case of attempted harm (negative belief, neutral outcome), as compared to each of the other conditions. (Post-hoc comparisons between attempted harm and the other three conditions revealed similar trends in the PC and LTPJ.) A whole-brain random effects group analysis also revealed a greater response uniquely in the RTPJ for attempted harm, contrasted with the all-neutral condition. One interpretation of these results is that moral condemnation depends more heavily on belief information in the absence of a negative outcome. That is, in the case of intentional harm (negative belief, negative outcome), the actor's causal role in bringing about an actual harm can contribute to moral condemnation. By contrast, in the case of attempted harm (negative belief, neutral outcome), moral condemnation rests solely on the agent's belief that his or her action will cause harm.

However, the response of the RTPJ (though not the PC or the LTPJ) at integration also showed an unexpected interaction with an additional variable: the order of belief and foreshadow information. In contrast to foreshadow-first trials, the RTPJ response at the time of integration of belief-first trials was significantly higher for neutral outcomes than negative outcomes, with no effect of belief valence. Consistent with this main effect, a whole-brain random effects group analysis revealed greater activation in the RTPJ for neutral versus negative outcomes. Moral judgments, by contrast, showed no interaction with order. One explanation for this effect is that participants "double-check" their previously encoded representation of the protagonist's beliefs more often, or more deeply, when belief information is presented early, relative to when belief information is presented immediately before the judgment. Future experiments will be necessary to test this hypothesis.

It is noteworthy that no aspect of the response in the RTPJ (or the PC or LTPJ) was determined simply by the truth or falsity of the beliefs, as has been suggested by recent work (Sommer et al., 2007). Consistent with our previous study (Young et al., 2007), the current results revealed a significantly above-baseline response in the RTPJ during integration in all eight moral conditions, half of which presented true beliefs; there was no main effect of truth at encoding or integration. We propose that the moral judgment task of the current study requires reasoning about beliefs, true or false. By contrast, the true belief trials of the "object transfer" task used in the previous research (Sommer et al., 2007) require participants to determine where an observer will look for an object that was "hidden" in full view of the observer. We suggest that this true belief task might not require belief reasoning at all; participants simply have to respond based on the true location of the object (Dennett, 1978). Robust recruitment of the RTPJ, PC, and LTPJ is observed for both true and false beliefs so long as belief reasoning is required by or relevant to the task.

We note that our interpretation of the current results is consistent with a specific role for the RTPJ, PC, and LTPJ in belief attribution. Both lesion and imaging studies implicate the RTPJ specifically, however, in another cognitive task: attentional reorienting in response to unexpected stimuli (Corbetta et al., 2000; Mitchell, 2007). Nevertheless, the RTPJ response in the current study is best understood as reflecting the processing of belief information, for two reasons. First, attentional reorienting cannot explain the highly selective functional response in the RTPJ. In the encoding phase, for example, belief and outcome information were equally frequent and equally expected, but the RTPJ responded selectively during sentences describing beliefs. Second, a recent study has found that the regions for belief attribution and exogenous attention are neighboring but distinct (Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, Saxe, personal communication). 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 and 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 as well as other regions implicated in theory of mind, i.e. the PC and LTPJ.

During both encoding and integration, regions in the MPFC showed a different functional profile. No region in the MPFC distinguished belief from foreshadow information during encoding. Therefore, even though the MPFC is routinely observed in the localizer task, there was no evidence for its specific role in the encoding of beliefs. During the integration phase, however, the dorsal MPFC was selective for the valence of the belief; its response was significantly higher for "negative" than for "neutral" beliefs. In other words, the dMPFC responded more when the protagonist thought that his or her action would cause harm, regardless of whether the action did cause harm.

There are two possible accounts for this effect: (1) the dMPFC is responsible for processing belief valence independent of the moral context and (2) the dMPFC is responsible for processing belief valence specifically for moral judgment. We favor the latter account for two reasons. First, belief valence was the dominant factor influencing participants' moral judgments in this study and other behavioral studies of adult moral judgments (Cushman, personal communication; Young et al., 2007). Second, the dMPFC was recruited differentially for negative over neutral beliefs not during encoding but, rather, only once subjects were able to make moral judgments of the agent's action, described only during the integration phase. These data suggest a role for the dMPFC in the evaluation of one kind of moral content, specifically, belief valence.

These results illuminate prior research suggesting a role for the MPFC in moral judgment (for a review, see Greene and Haidt, 2002;

Young and Koenigs, 2007). Previous research on the neural basis of moral judgment has focused largely on intentional harm; in all cases the protagonist both knows that his or her action will cause harm and does in fact cause harm by acting. Regions in the MPFC may therefore have been recruited for representing either actions that produce harmful *outcomes* or actions performed with harmful *intentions*. The current results suggest the latter: when subjects are presented with a description of the critical action, the dMPFC response is sensitive to whether the actor *thinks* he or she will cause harm by acting.

Conclusions

The current study reveals the neural basis of at least two distinct cognitive processes associated directly with theory of mind in moral judgment, the encoding and the integration of beliefs. Belief encoding is a stimulus-driven process: the response is based on whether the current stimulus contains belief information or not. Belief integration is a relatively stimulus-independent process: prior belief information is called upon and used in the service of moral judgment. A distinction between cognitive processes for encoding beliefs versus integrating beliefs into mature moral judgment is compatible with developmental research (Baird and Astington, 2004; Baird and Moses, 2001; Zelazo et al., 1996), and should be further investigated in developmental cognitive neuroscience. Differential development of function in theory of mind brain regions, including the RTPJ, PC, and LTPJ, may coincide with previously reported behavioral changes.

Both processes for belief attribution, though, appear to share a neural substrate in the temporo-parietal junction, bilaterally, and the precuneus. The medial prefrontal cortex, meanwhile, appears to be uniquely recruited for processing belief valence, a morally relevant feature of the action in the context of the task. These results may therefore inform future research probing the range of contexts both in and beyond the moral domain that depend on cognitive processes for encoding beliefs, integrating beliefs, and evaluating the valence of beliefs.

Acknowledgments

This project was supported by the National Center for Research Resources (grant P41RR14075), the MIND Institute, and the Athinoula A. Martinos Center for Biomedical Imaging. R.S. was supported by MIT and the John Merck Scholars program. L.Y. was supported by the NSF. Many thanks to Joshua Knobe, Fiery Cushman, and John Mikhail for comments on an earlier draft of this manuscript, Jonathan Scholz for technical assistance, and Alexandra Dickson and Neil Murthy for their help in data collection.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2008.01.057.

References

Aichorn, M., Perner, J., Kronblicher, M., Staffen, W., Ladurner, G., 2006. Do visual perspective tasks need theory of mind? NeuroImage 30, 1059–1068.

Baird, J.A., Astington, J.W., 2004. The role of mental state understanding in the development of moral cognition and moral action. New. Dir. Child. Adolesc. Dev. 103, 37–49.

Baird, J.A., Moses, L.J., 2001. Do preschoolers appreciate that identical actions may be motivated by different intentions? J. Cogn. Dev. 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. J. Cogn. Neurosci. 18, 803–817.

Ciaramidaro, A., Adenzato, M., Enrici, I., Erk, S., Pia, L., Bara, B.G., Walter, H., 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. Nat. Neurosci. 3, 292–297.

Cushman, F., Young, L., Hauser, M.D., 2006. The role of conscious reasoning and intuitions in moral judgment: testing three principles of harm. Psychol. Sci. 17, 1082–1089.

Darley, J.M., Zanna, M.P., 1982. Making moral judgment. Am. Sci. 70, 515–521.

Decety, J., Lamm, C., 2007. The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. The Neuroscientist 13, 580–593.

Dennett, D., 1978. Beliefs about beliefs. Behav. Brain Sci. 1, 568–570.

Fincham, F.D., Jaspers, J., 1979. Attribution of responsibility to the self and other in children and adults. J. Pers. Soc. Psychol. 37, 1589–1602.

Flavell, J.H., 1999. Cognitive development: children's knowledge about the mind. Ann. Rev. Psychol. 50, 21–45.

Fletcher, P.C., Happe, F., Frith, U., Baker, S.C., Dolan, R.J., Frackowiak, R.S.J., Frith, C.D., 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. J. Cogn. Neurosci. 19, 1803–1814.

Greene, J.D., Haidt, J., 2002. How (and where) does moral judgment work? Trends Cogn. Sci. 6, 517–523.

Grueneich, R., 1982. The development of children's integration rules for making moral judgments. Child Dev. 53, 887–894.

Hebble, P.W., 1971. Development of elementary school children's judgment of intent. Child Dev. 42, 583–588.

Karniol, R., 1978. Children's use of intention cues in evaluating behavior. Psychol. Bull. 85, 76–85.

Knobe, J., 2005. Theory of mind and moral cognition: exploring the connections. Trends Cogn. Sci. 9, 357–359.

Mikhail, J., 2007. Universal moral grammar: theory, evidence and the future. Trends Cogn. Sci. 11, 143–152.

Mitchell, J.P., 2007. Activity in right temporo-parietal junction is not selective for theory-of-mind. Cerebral Cortex 18, 262–271.

Nelson Le Gall, S.A., 1985. Motive outcome matching and outcome foreseeability—effects on attribution of intentionality and moral judgments. Dev. Psychol. 21, 332–337.

Nunez, M., Harris, P.L., 1998. Psychological and deontic concepts: separate domains or intimate connections. Mind Lang. 13, 153–170.

Onishi, K., Baillargeon, R., 2005. Do 15-month-old infants understand false beliefs. Science 308, 255–258.

Perner, J., Aichorn, M., Knronblicher, M., Staffen, W., Ladurner, G., 2006. Thinking of mental and other representations: the roles of left and right temporo-parietal junction. Soc. Neurosci. 1, 245–258.

Piaget, J., 1965/1932. The Moral Judgment of the Child. Free Press, New York.

Ruby, P., Decety, J., 2003. What you belief versus what you think they believe: a neuroimaging study of conceptual perspective-taking. Eur. J. Neurosci. 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. Psychol. Sci. 17, 692–699.

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

Shultz, T.R., Wright, K., Schleifer, M., 1986. Assignment of moral responsibility and punishment. Child Dev. 57, 177–184.

Siegel, M., Peterson, C.C., 1998. Preschoolers' understanding of lies and innocent and negligent mistakes. Dev. Psychol. 34, 332–341.

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., Maier, W., Shaw, N.J., Fink, G.R., Zilles, K., 2001. Mind reading: neural mechanisms of theory of mind and self-perspective. NeuroImage 14, 170–181.

Weiner, B., 1995. Judgments of Responsibility: a Foundation for a Theory of Social Conduct. New York, Guilford Press.

Wellman, H.M., Cross, D., Watson, J., 2001. Meta-analysis of theory-of-mind development: the truth about false belief. Child. Dev. 72, 655–684.

Wellman, H.M., Larkey, C., Somerville, S.C., 1979. Early development of moral criteria. Child Dev. 50, 869–873.

Young, L., Cushman, F., Hauser, M., Saxe, R., 2007. The neural basis of the interaction between theory of mind and moral judgment. Proc. Natl. Acad. Sci. 104, 8235–8240.

Young, L., Koenigs, M., 2007. Investigating emotion in moral cognition: A review of evidence from functional neuroimaging and neuropsychology. Br. Med. Bull. 84, 69–79.

Yuill, N., 1984. Young children's coordination of motive and outcome in judgements of satisfaction and morality. Br. J. Dev. Psychol. 2, 73–81.

Yuill, N., Perner, J., 1988. Intentionality and knowledge in children's judgments of actors responsibility and recipients emotional reaction. Dev. Psychol. 24, 358–365.

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