The neural basis of the interaction between theory of mind and moral judgment

Liane Young*†, Fiery Cushman*, Marc Hauser‡, and Rebecca Saxe§

*Department of Psychology and ‡Departments of Psychology, Organismic and Evolutionary Biology, and Biological Anthropology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138; and §Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139

Edited by Dale Purves, Duke University Medical Center, Durham, NC, and approved April 3, 2007 (received for review February 14, 2007)

Is the basis of criminality an act that causes harm, or an act undertaken with the belief that one will cause harm? The present study takes a cognitive neuroscience approach to investigating how information about an agent's beliefs and an action's consequences contribute to moral judgment. We build on prior developmental evidence showing that these factors contribute differentially to the young child's moral judgments coupled with neurobiological evidence suggesting a role for the right temporoparietal junction (RTPJ) in belief attribution. Participants read vignettes in a 2 × 2 design: protagonists produced either a negative or neutral outcome based on the belief that they were causing the negative outcome ("negative" belief) or the neutral outcome ("neutral" belief). The RTPJ showed significant activation above baseline for all four conditions but was modulated by an interaction between belief and outcome. Specifically, the RTPJ response was highest for cases of attempted harm, where protagonists were condemned for actions that they believed would cause harm to others, even though the harm did not occur. The results not only suggest a general role for belief attribution during moral judgment, but also add detail to our understanding of the interaction between these processes at both the neural and behavioral levels.

functional MRI | medial prefrontal cortex | morality | right temporoparietal junction | theory of mind

In the common law tradition, criminal conviction depends on both a harmful consequence (*actus reus*) and the intent to harm (*mens rea*) (1). In violation of this foundational legal principle, however, are crimes of attempt (2, 3). The incompetent criminal, for instance, who believes he has poisoned his victim but has instead administered only a harmless substance, can be convicted in a court of law. This poses a challenge to the philosophy of law: is the basis of criminality an act that causes harm, or an act undertaken with the belief that one will cause harm? We pursue a novel approach to this question based on the burgeoning research into the neurocognitive mechanisms of moral judgment, much of which has emphasized the role of multiple interacting systems (4–8). Specifically, we suggest that the apparent philosophical conflict between *actus reus* and crimes of attempt reflects the operation and integration of distinct mechanisms responsible for the processing of information about consequences and beliefs in the service of moral judgment.

From a developmental perspective, integrating information about mental states and outcomes presents a particular challenge for young children. When moral scenarios present conflicting information about the outcome of an action and the intention of the actor, young children's moral judgments and justifications are determined by the action's outcome rather than the actor's intention (9–13). For example, a person who intends to direct a traveler to the right location but accidentally misdirects him is judged by young children to be "naughtier" than a person who intends to misdirect a passerby but accidentally directs him to the right place (9). As children mature, they become progressively more likely to make the opposite judgment (11, 14–19). Although subsequent research has revealed that young children can use information about intentions to make moral distinctions when consequences are held constant between scenarios (14, 20–23), older children have consistently shown greater sensitivity to information about intentions. What develops then is not just "theory of mind," or the ability to represent the mental states of others, but the ability to integrate this information with information about consequences in the context of moral judgment (12, 13, 24, 25).

Developmental evidence thus suggests that mature moral judgments depend crucially on the cognitive processes responsible for representing and integrating information about beliefs and outcomes. Neuroimaging provides a useful tool for testing this hypothesis. To date, studies of the neural basis of moral judgment have focused primarily on emotional responses to intentional moral violations (6, 26–32). These studies suggest that regions in the medial prefrontal cortex (MPFC) are recruited for processing stimuli that visually depict or verbally describe moral violations. Convergent evidence from neuropsychological studies suggests that damage to these regions causes disturbances in moral behavior and moral reasoning (33–35). However, all of these investigations use vignettes featuring protagonists who act with the belief, stated or implied, that they will cause the outcome that they do cause, thereby confounding the dimensions of outcome and belief.

The neural basis of belief attribution has also been the topic of considerable research, revealing a consistent group of brain regions, including right temporoparietal junction (RTPJ), left temporoparietal junction (LTPJ), precuneus (PC), and MPFC (36–40). The RTPJ appears to be most selective for belief attribution (41–43); its response is high when subjects read stories that describe a character's true or false beliefs but low during stories containing other information about a character, including her appearance, cultural background, or even internal, subjective sensations (e.g., fatigue, hunger) that do not involve beliefs (e.g., representational content) (41). Previous neuroimaging studies of belief attribution, however, have neither included stimuli with a strong moral valence nor explored the interaction between the factors of belief and outcome in a moral context.

Author contributions: L.Y., F.C., M.H., and R.S. designed research; L.Y. and R.S. performed research; L.Y. and R.S. analyzed data; and L.Y., F.C., M.H., and R.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: RTPJ, right temporoparietal junction; PSC, percent signal change; LTPJ, left temporoparietal junction; PC, precuneus; MPFC, medial prefrontal cortex; dMPFC, dorsal MPFC; mMPFC, middle MPFC; vMPFC, ventral MPFC; ROI, region of interest; IPS, intraparietal sulcus; FEF, frontal eye field.

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

This article contains supporting information online at www.pnas.org/cgi/content/full/0701408104/DC1.

PSYCHOLOGY

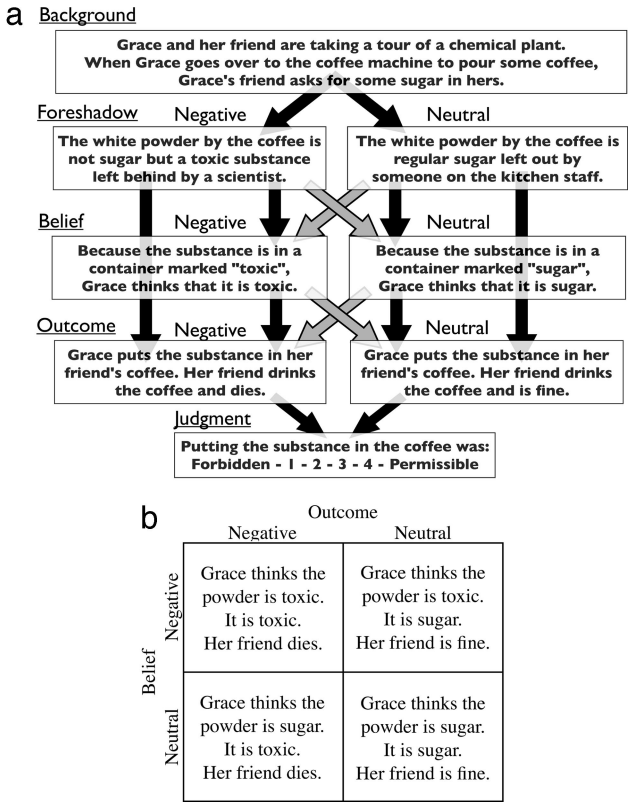

Fig. 1. Experimental stimuli and design. (*a*) Schematic representation of sample scenario. Light-colored arrows mark the combinations of ''Foreshadow'' and ''Belief'' for which the belief is false. ''Foreshadow'' information foreshadows whether the action will result in a negative or neutral outcome. ''Belief'' information states whether the protagonist holds a belief that she is in a negative situation and that action (or inaction) will result in a negative outcome (negative belief) or a belief that she is a neutral situation and that action will result in a neutral outcome (neutral belief). Sentences corresponding to each category were presented in 6-s blocks. (*b*) The combination of belief and outcome (as foreshadowed in ''Foreshadow'') yielded a 2 × 2 design and four conditions.

The current study used neuroimaging and behavioral methods to systematically investigate the interaction between belief attribution and moral judgment. Participants read vignettes in a 2 × 2 design (Fig. 1): protagonists produced either a negative outcome (someone's death) or a neutral outcome (no death) based on the belief that they were causing the negative outcome (''negative'' belief) or the neutral outcome (''neutral'' belief). In other words, a protagonist with a negative belief who produced a negative outcome did so knowingly, whereas a protagonist with a negative belief who produced a neutral outcome did so unknowingly based on a false belief (e.g., putting sugar in someone's coffee believing it to be poison). Participants judged the moral permissibility of the protagonist's action. This design allowed us to test distinct hypotheses about the recruitment of brain regions involved in belief attribution during moral judgment: (*i*) increased recruitment in the case that the protagonist believes he/she will cause harm (negative versus neutral beliefs), (*ii*) increased recruitment in the case that the protagonist's beliefs are incongruent with the outcome (false versus true beliefs), and (*iii*) the predicted interaction between belief and outcome such that recruitment is determined not just by the protagonist's beliefs but also by the consequences of his/her action. The experiment was also replicated in a second group of participants.

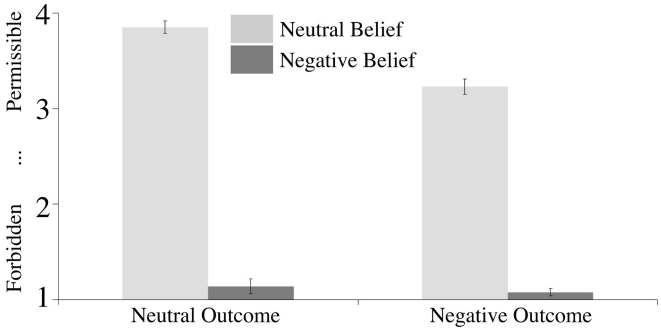

Fig. 2. Moral judgments given by subjects on a four-point scale (1, forbidden; 4, permissible). Error bars correspond to standard error.

Results and Discussion

Exp. 1 Behavioral Results. Subjects evaluated the moral status of protagonists' actions using four buttons on a scale from 1 for completely forbidden to 4 for completely permissible. A 2 × 2 repeated-measures ANOVA determined the influence of outcome (negative versus neutral) and belief (negative versus neutral) on judgments (Fig. 2). Actions performed by protagonists with negative beliefs were judged less permissible than when performed with neutral beliefs [negative, 1.1; neutral, 3.5; $F(1,9) = 712.4$; $P = 7.0 \times 10^{-10}$; partial $\eta^2 = 0.99$]. Subjects judged actions resulting in negative outcomes as less permissible than actions resulting in neutral outcomes [negative, 2.1; neutral, 2.5; $F(1,9) = 41.3$; $P = 1.2 \times 10^{-4}$; partial $\eta^2 = 0.82$].

The main effects were mediated by a significant interaction between belief and outcome [$F(1,9) = 21.2$; $P = 0.001$; partial $\eta^2 = 0.70$]. Specifically, post hoc Bonferroni's t tests revealed a significant difference between negative and neutral outcomes when the protagonist acted with a neutral belief [negative, 3.2; neutral, 3.9; $t(9) = -6.03$; adjusted $P = 3.8 \times 10^{-4}$] but no difference when the protagonist acted with a negative belief [negative, 1.1; neutral, 1.2; $t(9) = -1.83$; adjusted $P = 0.30$] (Fig. 2). When the protagonist believed she would not harm someone but in fact did (''unknowing harm''), her action was less permissible than a neutral action; when the protagonist believed she would harm someone but failed (attempted harm), her action was just as forbidden as if she had succeeded.

Reaction time data showed only an interaction between belief and outcome [$F(1,9) = 6.30$; $P = 0.03$] driven by marginally faster responses to the intentional harm condition (negative belief, negative outcome: 1.6 s) as compared with other conditions [all-neutral: 1.9 s, $t(9) = 0.3$; adjusted $P = 0.15$; attempted harm: 2.0 s, $t(9) = 0.3$; adjusted $P = 0.06$; unknowing harm: 2.0 s, $t(9) = 0.4$; adjusted $P = 0.03$].

Exp. 1 Regions of Interest (ROI) Analyses: Belief Attribution. 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, e.g., an outdated photograph (photo condition). A whole-brain random-effects analysis of the data replicated results of previous studies using the same task (39, 42), revealing higher BOLD response during belief, as compared with photo stories, in the RTPJ, dorsal MPFC (dMPFC), middle MPFC (mMPFC), ventral MPFC (vMPFC), PC, right temporal pole, and right anterior superior temporal sulcus ($P < 0.001$, uncorrected; $k > 10$). ROIs were identified in individual subjects (Table 1) at the same threshold: RTPJ (10/10 subjects), LTPJ (8/10), dMPFC (9/10), mMPFC (8/10), vMPFC (7/10), and PC (10/10).

The average percent signal change (PSC) from rest in each ROI was calculated for the third segment of each story (''be-

Table 1. Localizer experiment results

	Exp. 1						Exp. 2					
	Individual ROIs			Whole-brain contrast			Individual ROIs			Whole-brain contrast		
ROI	x	y	z	x	y	z	x	y	z	x	y	z
Belief > photo												
RTPJ	57	−56	24	56	−52	30	56	−56	22	56	−54	28
PC	1	−58	39	4	−62	38	−1	−58	39	0	−54	32
LTPJ	−49	−66	25	−42	−62	26	−50	−63	26	−52	−58	26
dMPFC	1	60	27	2	52	28	−2	58	29	2	60	28
mMPFC	−2	61	12	2	58	12	1	59	15	−4	56	8
vMPFC	1	57	−12	0	46	−2	1	55	−7	0	54	−8
Photo > rest												
Left IPS	−34	−59	52	−28	−66	58	−29	−60	51	−28	−58	50
Right IPS	33	−51	46	34	−48	46	33	−62	52	30	−62	48
Left FEF	−28	−1	64	−28	−2	62	−37	−5	64	−38	−4	64
Right FEF	38	3	62	38	2	64	34	−2	61	32	2	60

Average peak voxels for ROIs in Montreal Neurological Institute coordinates for Exps. 1 and 2. The ''Individual ROIs'' columns show the average of peak voxels from individual subjects' ROIs. The ''Whole-brain contrast'' columns show the peak voxel in the same regions in the whole-brain random-effects group analysis.

lief''), the first time at which all of the critical information for moral judgment (belief and outcome) was available. These responses were then analyzed by using a 2 × 2 repeated-measures ANOVA (Table 2). The RTPJ showed only a significant interaction between belief and outcome [$F(1,9) = 8.76$; $P = 0.02$] (Fig. 3). Planned comparisons revealed that the PSC was higher for negative belief than neutral belief in the case of neutral outcome [negative, 0.61; neutral, 0.27; $t(9) = 2.81$; $P = 0.02$] but was not significantly different for negative and neutral belief in the case of negative outcome [negative PSC, 0.25; neutral PSC, 0.28; $t(9) = 0.42$; $P = 0.68$). Post hoc Bonferroni's t tests revealed that the PSC for attempted harm was significantly greater than each of the other conditions [unknowing harm: $t(9) = 3.27$; adjusted $P = 0.02$; intentional harm: $t(9) = 4.09$; adjusted $P = 0.006$].

Analyses of the PSC averaged over the entire duration of the story revealed the same pattern of results [belief by outcome interaction: $F(1,9) = 18.53$; $P = 0.002$] [supporting information (SI) Fig. 4]. No effects were observed in other story segments independently: during the ''foreshadow'' segment of the story, the RTPJ did not discriminate between neutral (PSC: 0.37) and negative (PSC: 0.36) foreshadow; the ''outcome'' segment of the story did not reveal significant main effects or interactions.

Similar although subtly different patterns were found for other ROIs. An interaction between outcome and belief was also found in the PC [$F(1,9) = 12.05$; $P = 0.007$]; however, both of the paired contrasts were independently significant [neutral outcome, negative belief versus neutral belief: $t(9) = 3.38$; $P = 0.01$; negative outcome, negative belief versus neutral belief: $t(9) = 2.35$; $P = 0.04$]. In other words, the PSC in the PC was higher for false than true beliefs, but this effect was larger for negative than neutral beliefs. A similar pattern was found in the LTPJ [belief by outcome interaction: $F(1,7) = 19.54$; $P = 0.003$; neutral outcome, negative belief versus neutral belief: $t(7) = 4.12$; $P = 0.004$; negative outcome, negative belief versus neutral belief: $t(7) = 2.14$; $P = 0.06$]. The dMPFC showed a pattern similar to that of the RTPJ [belief by outcome interaction: $F(1,8) = 13.88$; $P = 0.01$; neutral outcome, negative belief versus neutral belief: $t(8) = 3.87$; $P = 0.01$; negative outcome, negative belief versus neutral belief: $t(8) = 0.15$; $P = 0.88$], except that the response for attempted harm was not significantly greater than the response for unknowing harm. Nonsignificant trends were found for mMPFC and vMPFC.

Exp. 1 ROI Analyses: Control. To test whether the observed effects were specific to brain regions implicated in belief attribution, we identified brain regions associated with general attention and response selection by comparing activity during the photo condition of the localizer experiment to baseline. A whole-brain random-effects analysis of the data replicated results of previous studies of brain regions associated with task performance (44, 45): higher PSC during photo stories, compared with rest, in the right and left intraparietal sulcus (IPS) and right and left frontal eye fields (FEF) ($P < 0.001$, uncorrected; $k > 10$). Each ROI was identified in every subject (Table 1) and analyzed by using the same 2 × 2 ANOVA as above. All of these regions showed a robust response in all conditions but no significant main effects or interactions. A 2 × 2 × 2 repeated-measures ANOVA for

Table 2. Belief attribution ROIs

	Mean PSC (belief, outcome)				Interaction of belief × outcome			
ROI	Neut, Neut	Neut, Neg	Neg, Neut	Neg, Neg	df	F	P value	Partial η^2
RTPJ	0.27	0.28	0.61	0.25	(1,9)	8.76	0.02	0.49
PC	0.13	0.31	0.4	0.08	(1,9)	12.05	0.01	0.57
LTPJ	0.31	0.56	0.59	0.36	(1,7)	19.54	0.003	0.74
dMPFC	−0.02	0.21	0.45	0.19	(1,8)	13.88	0.01	0.63
mMPFC	−0.45	−0.25	−0.08	−0.26	(1,7)	1.85	0.22	0.21
vMPFC	−0.08	−0.05	0.003	−0.005	(1,6)	0.03	0.87	0.005

Mean PSC in six ROIs during the belief segment of the moral scenarios. Four of these regions showed a significant interaction between negative (Neg) and neutral (Neut) belief and outcome information.

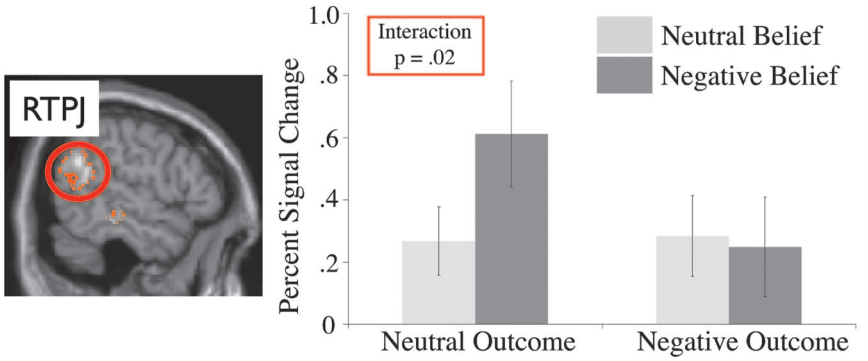

Fig. 3. PSC from rest in the RTPJ. (*Left*) Brain regions where the BOLD signal was higher for (nonmoral) stories about beliefs than (nonmoral) stories about physical representations ($n = 10$, random-effects analysis, $P < 0.001$ uncorrected). These data were used to define ROIs. (*Right*) The PSC in the RTPJ during the story segment when the protagonist's belief was stated (''Belief''). Error bars correspond to standard error.

every pair of regions that included one theory of mind region (e.g., RTPJ, PC, LTPJ, and dMPFC) and one task performance region (e.g., left IPS, right IPS, left FEF, and right FEF) revealed significant three-way interactions ($P < 0.05$) in every pair except the RTPJ and the left IPS, where the interaction approached significance [$F(1,9) = 4.28$; $P = 0.07$].

Exp. 1 Whole-Brain Analysis. Random-effects analyses of the whole brain were conducted for the main experiment ($P < 0.001$, uncorrected) (SI Table 3). A whole-brain analysis of the overall effect of belief (negative belief > neutral belief) revealed activation in the right anterior superior temporal sulcus and the dMPFC. A whole-brain analysis of the overall effect of outcome (negative outcome > neutral outcome) revealed no significant clusters. Whole-brain analyses of brain regions differentially activated for neutral > negative belief were conducted separately for (*i*) negative outcome and (*ii*) neutral outcome. The first contrast (unknowing harm > intentional harm) revealed activation in the right inferior parietal cortex, PC, right and left middle frontal gyrus, and right and left anterior cingulate sulcus. The second contrast yielded no significant clusters.

Exp. 2. The pattern of results observed in Exp. 1 was replicated in 17 new subjects. By using methods identical to Exp. 1, ROIs were identified in individual subjects (Table 1): RTPJ (15/17 subjects), LTPJ (16/17), dMPFC (14/17), mMPFC (12/17), vMPFC (10/17), and PC (17/17). In the RTPJ, a significant interaction between belief and outcome was observed during the same time interval [$F(1,14) = 10.14$; $P = 0.007$] (SI Fig. 5), and the PSC for attempted harm was significantly higher than for any other condition ($P < 0.05$). As in Exp. 1, similar patterns were found for other ROIs (SI Table 4). Combining the RTPJ data from both experiments in a $2 \times 2 \times 2$ ANOVA including gender as a factor, we found no significant main effect or interaction for gender. As in each experiment independently, the belief by outcome interaction in the whole sample ($n = 25$) was significant [$F(1,9) = 18.14$; $P = 3.0 \times 10^{-4}$].

General Discussion

At the broadest level, the results of the current study suggest that moral judgments depend on the cognitive processes mediated by the RTPJ, previously associated with belief attribution, and, to a lesser extent, the PC, LTPJ, and MPFC, which compose a network of brain regions implicated in theory of mind. Specifically, the results reveal significantly above-baseline activation of the RTPJ for all four conditions (intentional harm, attempted harm, unknowing harm, and all-neutral), highlighting the role of belief attribution during moral judgment. Importantly, however, brain regions involved in belief attribution were not recruited indiscriminately across conditions. In particular, we found a selective increase in the response for the case of attempted harm, in which the protagonist believed that he would harm someone but in fact did not. The differential neural response between experimental conditions suggests an unequal contribution of belief attribution to moral judgment depending not only on what the protagonist believes, as might be expected, but also on the consequences of the protagonist's behavior. This result offers a new perspective on the integration of information about beliefs and consequences in moral judgment, the focus of our discussion.

The behavioral data suggest that, across conditions, moral judgment is determined primarily by belief information, consistent with the robust RTPJ response for all four conditions. An interesting asymmetry emerged, however, for cases in which belief and outcome information were in conflict, as in situations of attempted harm and unknowing harm. We found that subjects' moral judgments were determined solely by belief in the case of attempted harm but not unknowing harm. That is, attempted harm (e.g., putting sugar in a friend's coffee believing it to be poison) was judged fully forbidden, just as though the protagonist had successfully produced the negative outcome of the friend's death. By contrast, moral judgment of unknowing harm appeared to depend on both the outcome of the action and on the belief state of the actor. Unknowing harm (e.g., putting poison in a friend's coffee believing it to be sugar) was not judged fully permissible, as compared with the all-neutral condition, in which the protagonist held a neutral belief and produced a neutral outcome.

This interpretation of the behavioral data is consistent with the activation profile of the RTPJ, as suggested by the ROI analyses. Although we observed a robust BOLD response in the RTPJ for all four conditions, we observed an interaction between the consequences of the protagonist's action and the contents of the protagonist's belief: the attempted harm condition elicited the highest response. The RTPJ response was not determined simply by false (versus true) beliefs or by negative (versus neutral) beliefs. Instead, the RTPJ response was selectively enhanced when subjects used information about a protagonist's beliefs to condemn the protagonist despite his failing to cause harm. The same basic activation pattern, although weaker and less selective, was observed in other brain regions implicated in theory of mind. Importantly, if the activation in these regions reflected only the process of determining and representing the protagonist's belief, no interaction would be expected. These results also suggest an asymmetry in the cognitive processes that give rise to the moral condemnation of successful and unsuccessful attempts to harm. The condemnation of successful crimes relies less on belief attribution, presumably because moral condemnation can rest on causal responsibility for an actual

harm. By contrast, the condemnation of failed attempts relies heavily on belief attribution.

The pattern of brain activation linked to belief attribution cannot be attributed to an increase in general attention or effort for the condition of attempted harm. First, reaction time data revealed an interaction driven by faster responses to the intentional harm condition, as compared with any other condition; there was no difference between any of the other conditions. These results make intuitive sense: moral judgments are relatively rapid when harm is done and all possible pieces of information are consistent, as in the case of intentional harm. By contrast, the response of the RTPJ was highest for attempted harm. The RTPJ response therefore could not be explained by increased, or decreased, time-on-task. Second, brain regions implicated in attention and response selection (e.g., IPS and FEF) did not discriminate between the moral judgment story conditions (although these regions were recruited robustly for all conditions).

The current results also reveal an asymmetry between moral judgments of incompetent criminals (whose false beliefs prevent intended harm from occurring) and unlucky innocents (whose false beliefs lead them to cause unintended harms). Judgments of incompetent criminals were harsh, made on the basis of beliefs alone, and associated with enhanced recruitment of circuitry involved in belief attribution. By contrast, unlucky innocents were not entirely exculpated for causing harm on the basis of their false beliefs. Instead of showing an increased response in brain regions associated with belief attribution, whole-brain analyses revealed recruitment of brain regions associated with cognitive conflict: right inferior parietal cortex, PC, bilateral middle frontal gyrus, and bilateral anterior cingulate sulcus. All of these regions have been implicated in cognitive conflict associated with moral dilemmas (6), specifically where subjects endorse emotionally salient harmful acts to prevent greater harm. Here subjects had to override judgments against harm in favor of utilitarian considerations (e.g., the greatest good for the greatest number). Analogously, in the context of unknowing harm, subjects may partially override judgments against harm to exculpate agents on the basis of their false beliefs. Moral judgment may therefore represent the product of two distinct and at times competing processes, one responsible for representing harmful outcomes and another for representing beliefs and intentions (F.C., unpublished data).

The interpretation we offer here is compatible with the significant developmental literature showing that young children's moral judgments are determined primarily by information about the outcomes of actions rather than the intentions of the actor (13, 16, 19). This pattern, in conjunction with evidence showing that young children lack a mature theory of mind (47, 48), indicates a dissociation between two processes important for mature moral judgment. With the development of theory of mind, young children are progressively able to integrate belief information in moral judgment (12). We suggest that a late-developing process for representing mental states together with an early-developing process responsible for representing harmful consequences contribute to moral judgment in mature adults, and, in some cases, these processes may interact competitively. Research into developmental disorders such as autism, characterized by deficits in theory of mind, should provide further insight into the relationship between theory of mind and moral judgment (49, 50).

In summary, the results of the present study demonstrate systematically different patterns of reliance on mental state attribution for the cases considered, patterns that cannot be accounted for by the presence or absence of false or negative beliefs. Future investigations will be necessary to explore the neural basis of intuitive moral and legal judgments in a range of related cases that depend on both belief attribution and harm detection (51). Other impossible attempts (e.g., voodoo) as well as other crimes that occur mainly in the mind (e.g., conspiracy) should prove to be particularly interesting, as will crimes where the true beliefs and intentions of the actor are not given but must be inferred. Investigating the neural processes that support belief attribution across these different contexts will prove informative for cognitive models of theory of mind and moral reasoning.

Methods

Exp. 1. Ten 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 Internal Review Boards at Massachusetts General Hospital and the Massachusetts Institute of Technology. Subjects were scanned at 3 T (at the Massachusetts General Hospital scanning facility in Charlestown, MA) by using 26 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 four variations of 12 scenarios for a total of 48 stories with an average of 86 words per story (see *SI Text* for full text of scenarios). A 2 × 2 design was used for each scenario: protagonists (*i*) produced either a negative outcome or a neutral outcome based on (*ii*) the belief that they were causing a negative outcome or a neutral outcome. Stories were presented in four cumulative segments, each presented for 6 s, for a total presentation time of 24 s per story: (*i*) background, information to set the scene (identical across all conditions); (*ii*) foreshadow, information foreshadowing the outcome (negative or neutral); (*iii*) belief, information stating the protagonist's belief about the situation (negative or neutral); (*iv*) outcome, information about the protagonist's action and resulting outcome.

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), the protagonist's action or inaction results in someone's death. Each possible belief was true for one outcome and false for the other outcome. Presentation time was rapid (marginally longer than the mean reading time required by subjects in a self-paced pilot version) to motivate subjects to read the stimuli as they were presented. Stories were then removed from the screen and replaced with a question about the moral nature of the action on a scale of 1 (forbidden) to 4 (permissible) using a button press. The question remained on the screen for 4 s.

Subjects saw two variations of each scenario, for a total of 24 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. When a scenario was repeated, it contained a different protagonist (i.e., first name), belief, and outcome from the first presentation. Six stories were presented in each 4-min, 24-s run; the total experiment, involving four runs, lasted 18 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.

In the same scan session, subjects participated in four runs of a localizer experiment, contrasting stories that required inferences about a character's beliefs with stories that required inferences about a physical representation, i.e., a photo that has become outdated. Stimuli and story presentation were exactly as described for Saxe and Kanwisher's experiment 2 (39).

Exp. 2. Seventeen new subjects (Harvard College undergraduates, aged 18–22 years, six women) meeting the same criteria identified in Exp. 1 participated in a second functional MRI experiment. Scanning was conducted at the Massachusetts Institute of Technology; otherwise, all scan parameters were identical. Because of technical limitations, subjects used only three buttons to respond. An expanded set of stories was used, including versions of the 24 original scenarios and 24 new scenarios with the same structure; each scenario was presented only once. Subjects saw 24/48 stories in the original four conditions (belief by outcome). Four other conditions were presented that are not considered here. All analyses followed the same procedures, as described below.

Functional MRI Analysis. MRI data were analyzed by using SPM2 (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 template). Data were then smoothed by 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 by 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 order and timing of the four story components were 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 two sets of tailored ROI analyses were conducted. First, 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, dMPFC, mMPFC, vMPFC, and PC. Second, four ROIs were defined for each subject individually based on a whole-brain analysis of another localizer contrast and defined as contiguous voxels that were significantly more active ($P < 0.001$, uncorrected) while the subject read photo stories, as compared with baseline: right IPS, left IPS, right FEF, and left FEF. All peak voxels are reported in Montreal Neurological Institute coordinates.

The responses of these ROIs were then measured while subjects read stories from the current experiment. Within the ROI, the average 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.

We thank Joshua Knobe, Ralph Adolphs, Laura Scholz, Walter Sinnott-Armstrong, and Edouard Machery for comments on an earlier draft. We thank Nancy Kanwisher for generous support and Jonathan Scholz for technical assistance. This work was made possible by the Athinoula A. Martinos Imaging Center and the Guggenheim Foundation.

1. Hart HLA (1968) *Punishment and Responsibility* (Oxford Univ Press, Oxford).
2. Brown DE (1991) *Human Universals* (McGraw–Hill, New York).
3. Fletcher G (1998) *Basic Concepts of Criminal Law* (Oxford Univ Press, Oxford).
4. Hauser MD (2006) *Moral Minds: How Nature Designed a Universal Sense of Right and Wrong* (HarperCollins, New York).
5. Knobe J (2005) *Trends Cognit Sci* 9:357–359.
6. Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD (2004) *Neuron* 44:389–400.
7. Haidt J (2001) *Psychol Rev* 108:814–834.
8. Cushman FA, Young L, Hauser MD (2006) *Psychol Sci* 17:1082–1089.
9. Piaget J (1965/1932) *The Moral Judgment of the Child* (Free Press, New York).
10. Hebble PW (1971) *Child Dev* 42:583–588.
11. Shultz TR, Wright K, Schleifer M (1986) *Child Dev* 57:177–184.
12. Yuill N, Perner J (1988) *Dev Psychol* 24:358–365.
13. Zelazo PD, Helwig CC, Lau A (1996) *Child Dev* 67:2478–2492.
14. Karniol R (1978) *Psychol Bull* 85:76–85.
15. Fincham FD, Jaspers J (1979) *J Pers Soc Psychol* 37:1589–1602.
16. Baird JA, Astington JW (2004) *New Directions Child Adolescent Dev* 103:37–49.
17. Darley JM, Zanna MP (1982) *Am Sci* 70:515–521.
18. Yuill N (1984) *Br J Dev Psychol* 2:73–81.
19. Baird JA, Moses LJ (2001) *J Cognition Dev* 2:413–448.
20. Nelson Le Gall SA (1985) *Dev Psychol* 21:332–337.
21. Wellman HM, Larkey C, Somerville SC (1979) *Child Dev* 50:869–873.
22. Nunez M, Harris PL (1998) *Mind Lang* 13:153–170.
23. Siegel M, Peterson CC (1998) *Dev Psychol* 34:332–341.
24. Grueneich R (1982) *Child Dev* 53:887–894.
25. Imamoglu EO (1975) *Child Dev* 46:39–45.
26. Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) *Science* 293:2105–2108.
27. Borg JS, Hynes C, Van Horn J, Grafton S, Sinnott-Armstrong W (2006) *J Cognit Neurosci* 18:803–817.
28. Moll J, de Oliveira-Souza R, Bramati IE, Grafman J (2002) *NeuroImage* 16:696–703.
29. Moll J, de Oliveira-Souza R, Eslinger PJ, Bramati IE, Mourao-Miranda J, Andreiulo PA, Pessoa L (2002) *J Neurosci* 22:2730–2736.
30. Moll J, de Oliveira-Souza R, Moll FT, Ignacio FA, Bramati IE, Caparelli-Daquer EM, Eslinger PJ (2005) *J Cognit Behav Neurol* 18:68–78.
31. Heekeren HR, Wartenburger I, Schmidt H, Schwintowski HP, Villringer A (2003) *NeuroReport* 14:1215–1219.
32. Luo Q, Nakic M, Wheatley T, Richell R, Martin A, Blair RJ (2006) *NeuroImage* 30:1449–1457.
33. Anderson SW, Bechara A, Damasio H, Tranel D, Damasio AR (1999) *Nat Neurosci* 2:1032–1037.
34. Eslinger PJ, Grattan LM, Damasio AR (1992) *Arch Neurol* 49:764–769.
35. Koenigs M, Young L, Adolphs R, Tranel D, Cushman F, Hauser M, Damasio A (2007) *Nature* 446:908–911.
36. Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, Frackowiak RSJ, Frith CD (1995) *Cognition* 57:109–128.
37. Gallagher HL, Happe F, Brunswick N, Fletcher PC, Frith U, Frith CD (2000) *Neuropsychologia* 38:11–21.
38. Ruby P, Decety J (2003) *Eur J Neurosci* 17:2475–2480.
39. Saxe R, Kanwisher N (2003) *NeuroImage* 19:1835–1842.
40. Vogeley K, Bussfield P, Newen A, Herrmann S, Happe F, Falkai P, Maier W, Shaw NJ, Fink GR, Zilles K (2001) *NeuroImage* 14:170–181.
41. Saxe R, Powell L (2006)) *Psychol Sci* 17:692–699.
42. Saxe R, Wexler A (2005) *Neuropsychologia* 43:1391–1399.
43. Aichhorn M, Perner J, Kronbichler M, Staffen W, Ladurner D (2006) *NeuroImage* 30:1059–1068.
44. Saxe R, Schulz L, Jiang L (2006) *Soc Neurosci* 1:284–298.
45. Jiang Y, Kanwisher N (2003) *J Cognit Neurosci* 15:1095–1110.
46. Grant K, Boucher J, Riggs KJ, Grayson A (2005) *Autism* 9:317–331.
47. Wimmer H, Perner J (1983) *Cognition* 13:103–128.
48. Wellman HM, Cross D, Watson J (2001) *Children Dev* 72:655–684.
49. Blair RJ (1996) *J Autism Dev Disorders* 26:571–579.
50. Leslie A, Mallon R, Dicorcia J (2006) *Soc Neurosci* 1:270–283.
51. Mikhail J (2007) *Trends Cognit Sci* 11:143–152.