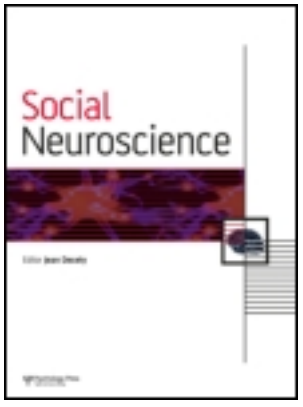

Social Neuroscience
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/psns20

Neural evidence for "intuitive prosecution": The use of mental state information for negative moral verdicts

Liane Young [a], Jonathan Scholz [a] & Rebecca Saxe [a]
[a] Massachusetts Institute of Technology, Cambridge, MA, USA

Available online: 25 Jan 2011

To cite this article: Liane Young, Jonathan Scholz & Rebecca Saxe (2011): Neural evidence for "intuitive prosecution": The use of mental state information for negative moral verdicts, Social Neuroscience, 6:3, 302-315

To link to this article: http://dx.doi.org/10.1080/17470919.2010.529712

SOCIAL NEUROSCIENCE, 2011, 6 (3), 302–315

Psychology Press
Taylor & Francis Group

Neural evidence for "intuitive prosecution": The use of mental state information for negative moral verdicts

Liane Young, Jonathan Scholz, and Rebecca Saxe
Massachusetts Institute of Technology, Cambridge, MA, USA

Moral judgment depends critically on theory of mind (ToM), reasoning about mental states such as beliefs and intentions. People assign blame for failed attempts to harm and offer forgiveness in the case of accidents. Here we use fMRI to investigate the role of ToM in moral judgment of harmful vs. helpful actions. Is ToM deployed differently for judgments of blame vs. praise? Participants evaluated agents who produced a harmful, helpful, or neutral outcome, based on a harmful, helpful, or neutral intention; participants made blame and praise judgments. In the right temporo-parietal junction (right TPJ), and, to a lesser extent, the left TPJ and medial prefrontal cortex, the neural response reflected an interaction between belief and outcome factors, for both blame and praise judgments: The response in these regions was highest when participants delivered a *negative* moral judgment, i.e., assigned blame or withheld praise, based solely on the agent's intent (attempted harm, accidental help). These results show enhanced attention to mental states for negative moral verdicts based exclusively on mental state information.

Keywords: Morality; Blame; Praise; Theory of mind; Temporo-parietal junction.

INTRODUCTION

Many recent studies have targeted the cognitive processes and neural substrates that support moral judgment (Cushman, Young, & Hauser, 2006; Gazzaniga, 2005; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001; Inbar, Pizarro, Knobe, & Bloom, 2009; Mikhail, 2007; Moll et al., 2005; Wheatley & Haidt, 2005). The majority of these studies focus on participants' negative evaluations of moral violations; for instance, hitting people with trolleys, breaking promises, distributing resources unfairly, and eating dead pets (Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006; Cushman, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Hsu, Anen, & Quartz, 2008). Moral judgments across these cases reflect a multitude of cognitive processes, including emotional responses to bad behavior and its effects (Harenski & Hamaan, 2006; Heekeren, Wartenburger, Schmidt, Schwintowski, & Villringer, 2003; Young et al., 2010), as well as representations of the agent's mind, including his or her beliefs and intentions, i.e. "theory of mind" (ToM) (Borg et al., 2006; Young, Cushman, Hauser, & Saxe, 2007). Moral psychology, however, encompasses not just negative evaluation but also positive evaluation, which has received less attention so far. The positive psychology movement (Seligman & Csikszentmihalyi, 2000) has led some researchers to study positive moral emotions (Haidt, 2003) and the neural signatures of cooperative behavior (de Quervain et al., 2004; Moll et al., 2006; Rilling et al., 2002) as well as subjective responses to moral virtues (Takahashi et al., 2008). These studies have focused primarily on the distinctiveness of positive emotions and their neural substrates.

Correspondence should be addressed to: Liane Young, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, USA. E-mail: liane.young@bc.edu

This project was supported by the Athinoula A. Martinos Center for Biomedical Imaging. The authors were supported by the Simons Foundation, the NSF, and the John Merck Scholars program. Many thanks to Fiery Cushman for helpful comments on an earlier draft and to Riva Nathans and Allan Mintz for help with stimulus construction and data collection.

www.psypress.com/socialneuroscience DOI: 10.1080/17470919.2010.529712

The current study seeks to extend this tradition by taking a different approach. Here we focus on one of the many cognitive processes implicated in moral judgment—theory of mind—for evaluating not only harmful but also helpful actions. Prior behavioral work suggests that theory of mind may play different roles in moral blame vs. praise. First, people assign less blame for impulsive as compared to deliberate harms (e.g., crimes of passion vs. premeditated crimes) but do not distinguish between impulsive and deliberate helpful actions (Pizarro, Uhlmann, & Salovey, 2003). Second, people judge actions with negative side-effects to be more intentional (e.g., supporting a profitable policy that also harms the environment) than actions with positive side effects (Knobe, 2005). Third, people rely on different kinds of mental states, in the case of side-effects; blame is based relatively more on the agent's belief (e.g., that harm will be done), and praise on the agent's desire (e.g., to be helpful; F. Cushman, personal communication).

The current study uses functional magnetic resonance imaging (fMRI) to investigate the role of ToM for moral judgments of blame and praise. At the broadest level, we aim to investigate whether brain regions that support ToM for non-moral judgments (e.g., behavior prediction and explanation) are differentially recruited for evaluating harmful and helpful actions, and whether, within this neural network, the same brain regions are recruited for blame and praise.

This study therefore builds on prior fMRI investigations into ToM in non-moral contexts. These prior studies show consistent neural activation for the processing of verbal and visual stimuli that depict mental states: the medial prefrontal cortex (MPFC), right and left temporo-parietal junction (RTPJ, LTPJ), and precuneus (den Ouden, Frith, Frith, & Blakemore, 2005; Fletcher et al., 1995; Frith & Frith, 2003; Gallagher et al., 2000; Ruby & Decety, 2003; Saxe & Kanwisher, 2003; Vogeley et al., 2001). Of these regions, the RTPJ has been shown to be particularly selective for processing mental states with representational content such as thoughts and beliefs (Aichhorn, Perner, Kronbichler, Staffen, & Ladurner, 2006; Ciaramidaro et al., 2007; Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Saxe & Wexler, 2005). For example, the response in the RTPJ is high when participants read stories that describe a person's beliefs, true or false, but low during other socially salient stories describing, for example, a person's physical appearance, cultural background, or even internal subjective sensations that lack representational content, i.e. hunger or fatigue (Saxe & Powell, 2006). Typically, the LTPJ shows a similar response profile; however, recent work suggests the LTPJ may play a more general role in representation selection, regardless of the content of the representation (Perner et al., 2006). More specifically, Perner and colleagues found that the LTPJ is activated not only by false beliefs but also by false signs, indicating that the LTPJ may be responsible for processing generic perspective differences in both the mental and the nonmental domain (Perner et al., 2006). By contrast, the RTPJ was activated only for false beliefs.

The critical role of these brain regions, including the RTPJ, for evaluating harmful actions has also been the topic of recent research, using transcranial magnetic stimulation (TMS) (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010) and fMRI (Young et al., 2007). For example, the same regions for ToM in nonmoral contexts were recruited when participants read explicit statements of agents' beliefs about whether or not they would cause harm (e.g., "Grace thinks the powder is poison") and then judged the moral permissibility of the action (e.g., "Grace puts the powder in her friend's coffee") (Young & Saxe, 2008). During the moral judgment, the RTPJ showed not only a robust response but also an interaction between belief and outcome (Young et al., 2007): The RTPJ response was significantly higher for failed attempts to harm (negative belief/intent, neutral outcome), as compared to all other conditions, including the other false belief condition, i.e., accidental harm (neutral belief/intent, negative outcome). In general, this interaction suggests that the RTPJ is involved not only in the initial encoding of the explicitly stated belief, as well as perhaps the inferred intention, but also in the integration of the belief with the outcome for moral judgment. Moreover, the precise pattern of activation (i.e., high response for attempted but not accidental harms) shows that the RTPJ does not simply respond to false beliefs, which are incompatible with the actual outcomes. Convergent TMS evidence suggests that temporarily disrupting RTPJ activity using online and offline TMS has the most pronounced effect on moral judgments of attempted harms as well, biasing participants to judge attempted harms more leniently, based on the neutral outcome rather than the negative intent (Young et al., 2010a).

The functional profile observed in the RTPJ then presents a puzzle. Why is the RTPJ most robustly recruited during moral judgments of attempted harms? One interpretation is that the enhanced RTPJ activation reflects greater attention to or deeper encoding of mental states when moral judgments depend primarily on mental states. Moral condemnation in the absence of an actual harm (e.g., attempted harm) must depend heavily on the agent's belief or

intention. By contrast, in the case of intentional harm, the actor's causal role in bringing about an actual harm might additionally contribute to moral condemnation (Cushman, 2008). However, a problem for this interpretation is the lower response to accidental harms. Forgiving or exculpating an agent for causing harm accidentally, based on a false belief (Young & Saxe, 2009b), must also depend heavily on a representation of the agent's mental state, specifically the false belief. The pattern of results thus suggests an amended view: the neural processes for mental state reasoning are most robustly recruited when a *negative* moral judgment depends on the agent's belief or intent. In other words, moral judgment and mental state reasoning may interact such that (1) mental states (in this case, beliefs or inferred intentions) are weighed more heavily when they form the predominant basis of moral judgment (e.g., when the belief/intent conflicts with the outcome), and (2) mental states are weighed more heavily for negative (as opposed to neutral or positive) moral judgments. These two influences may underlie the pattern of neural activation. We'll call this hypothesis the "intuitive prosecutor hypothesis" whereby participants attend especially to evidence (here, mental state evidence) that supports a relatively negative moral verdict; in other words, it shifts moral judgments downward, assigning blame in the absence of a negative outcome, or withholding praise in the presence of a positive outcome.

On the other hand, the interaction observed in the RTPJ could also be explained by an alternative account. On this view, which we'll call the "goal incompletion hypothesis" (R. Baillargeon, personal communication), the enhanced RTPJ activation reflects the processing of a salient goal (e.g., trying to poison a friend) that the agent fails to complete, as in the case of a failed murder attempt. The response is thus low for intentional harms, because the agent successfully completes the salient goal, and low for accidental harms, because the original goal of the action, which the agent failed to complete, was not especially salient (e.g., sweetening a friend's coffee). On the goal incompletion hypothesis then, participants attend especially to salient mental states, such as murderous desires, that don't amount to murder in the end.

The current paper both (1) investigates the neural processes that support ToM for blame vs. praise, and (2) tests the intuitive prosecutor vs. goal incompletion hypotheses. Participants read modified versions of the harm scenarios used in our previous research as well as new "help" scenarios, both in a 2×2 design: Protagonists produced a valenced (harmful or helpful) outcome or neutral outcome, based on a valenced or neutral intent. Participants made judgments of moral blame (for harm scenarios) and moral praise (for help scenarios). In general, we tested whether the same ToM brain regions would be recruited for both kinds of moral judgments. More specifically, using scenarios featuring positive goals (e.g., helping other people) allowed us to test our two hypotheses. When agents attempt to help others but fail ("attempted help"), their goals are salient but incomplete. When agents end up helping others accidentally, based on false beliefs and no intention to help ("accidental help"), then a relatively *negative* moral judgment (withholding praise) is based on the belief or intention. (We note that a "negative moral judgment" in the case of the help scenarios may still involve praise, only low levels of praise.) On the goal incompletion hypothesis, participants reason more about any salient incomplete goal; therefore, the RTPJ response should be high for attempted help, just like attempted harm, and lower for accidental help. On the intuitive prosecutor hypothesis, participants reason more about beliefs and intentions that support negative moral judgments; therefore, the RTPJ response should be high for accidental help (low praise), just like attempted harm (high blame), and lower for attempted help.

METHODS

Seventeen right-handed subjects (aged 18–22 years, 10 women) participated in the study for payment. Behavioral data were collected but later lost from the first five subjects; behavioral analyses therefore reflect data from 12 subjects (eight women) (see "Supplementary information"). 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 Board at MIT. Subjects were scanned at 3 T (at the MIT scanning facility in Cambridge, MA) 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 two sets of scenarios: (1) four variations (conditions) of 24 harm scenarios and (2) four variations of 24 help scenarios for a total of 192 stories (see Figure 1 for sample scenarios and www.mit.edu/~lyoung/files for full text). For *harm* scenarios: (i) Agents produced either a negative outcome (harm to others) or a neutral outcome (no harm), and (ii) agents believed they were causing a negative outcome ("negative" belief/intent) or a neutral outcome ("neutral" belief/intent). For *help* scenarios:

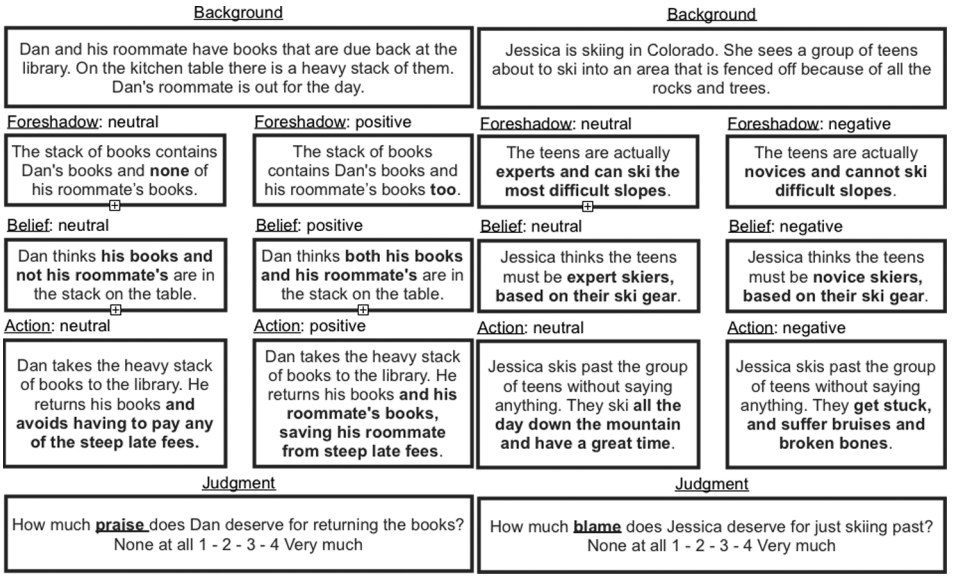

Figure 1. Schematic representation of sample help (left) and harm (right) scenarios. Changes across conditions are shown in bold text. "Background" information sets the scene. "Foreshadow" information foreshadows whether the action will result in a positive/negative or neutral outcome. "Belief" information states whether the protagonist holds a belief that she is in a positive/negative situation and that action will result in a positive/negative outcome (positive/negative belief) or a belief that she is in a neutral situation and that action will result in a neutral outcome (neutral belief). "Action" information describes the action and its outcome. Subjects made praise/blame judgments of protagonists' actions. Sentences corresponding to each category were presented in 6 s segments.

(i) Agents produced either a positive outcome (help to others) or a neutral outcome (no help), and (ii) agents believed they were causing a positive outcome ("positive" belief/intent) or a neutral outcome ("neutral" belief/intent). Helpful outcomes included benefits to others. Harmful outcomes included injuries to others. Word count was matched across harm conditions and help conditions (see "Supplementary information"). Stories were in four cumulative segments, each presented for 6 s, for a total presentation time of 24 s per story (see Figure 2 for timeline of a single trial):

1. *background*: information to set the scene (identical across conditions)
2. *foreshadow*: information foreshadowing outcome (valenced or neutral)
3. *belief*: the agent's belief about the situation (valenced or neutral)
4. *action and outcome*: the agent's action and actual outcome (valenced or neutral).

We note that while the stimuli explicitly specified the agent's belief about whether he or she would harm or help another person, participants could also infer the agent's intention with respect to the action and outcome. Pilot behavioral data suggest that the current stimuli support assumptions about the agents' desires and intentions, i.e. if Grace *thought* the powder was poison, she probably *wanted* to poison her friend.

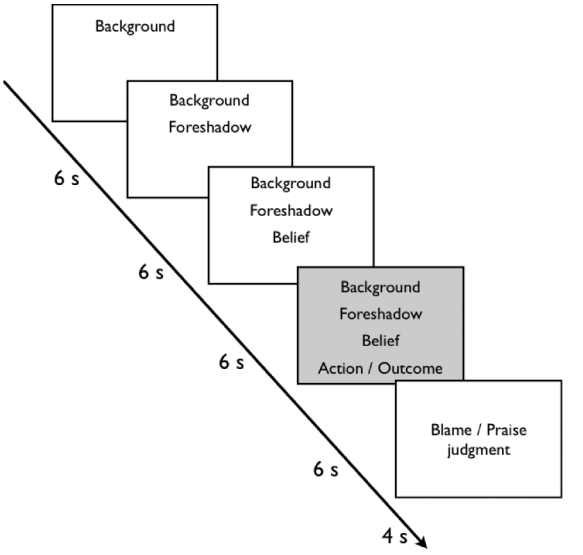

Figure 2. Schematic representation of a single moral judgment trial. Stories were presented in four cumulative segments, each presented for 6 s, for a total presentation time of 24 s per story. The story was then removed, and replaced by a question, for 4 s, concerning how much moral blame (for harm scenarios) or praise (for help) the protagonist deserves for acting, from none (1) to a lot (4). During the critical segment (shaded), all morally relevant information was made available for participants to use in moral judgment.

Each version of the belief was true for one outcome and false for the other outcome (e.g., the negative belief was true if the story ended with the negative

outcome and false if the story ended with the neutral outcome). After 24 s, the story was removed, and replaced by a question concerning how much moral blame (for harm scenarios) or praise (for help scenarios) the protagonist deserves for his or her action, from none (1) to a lot (4), using a button press. The question was 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, while 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.

In the same scan session, subjects participated in four runs of a ToM localizer task (Saxe & Kanwisher, 2003), contrasting stories requiring inferences about false beliefs with control stories, matched for linguistic complexity and logical structure, requiring inferences about "false" physical representations, i.e., a photograph or map that had become outdated. Stimuli and story presentation for the ToM localizer task were exactly as described in Saxe & Kanwisher (2003), Experiment 2.

FMRI analysis

MRI data were analyzed using SPM2 (www.fil.ion.ucl.ac.uk/spm) and custom software. Each subject's data were motion corrected and normalized onto a common brain space (Montreal Neurological Institute, MNI, template). Data were 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 to estimate the hemodynamic response for each condition. An event was defined as a single story, and the event onset was defined by the onset of text on screen. The timing of 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 regions of interest (ROIs; see below).

Both random effects whole-brain analyses (over the entire time course) and tailored ROI analyses were conducted. Six ROIs were defined for each subject individually based on a whole-brain analysis of the independent localizer experiment, and defined as contiguous voxels that were significantly more active ($p < 0.001$, uncorrected) (Saxe, Brett, & Kanwisher, 2006) while the subject read belief stories, as compared with photograph stories. All peak voxels are reported in MNI coordinates (Table 1).

TABLE 1
Localizer experiment results

	Individual ROIs			*Whole-brain contrast*		
ROI	*x*	*y*	*z*	*x*	*y*	*z*
RTPJ	58	−55	22	56	−54	24
PC	0	−57	40	−2	−56	46
LTPJ	−52	−59	26	−50	−54	26
dMPFC	1	56	38	−2	58	32
mMPFC	1	60	17	−6	50	24
vMPFC	1	58	−12	−2	54	−14

Notes: Average peak voxels for ROIs in MNI 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.

The responses of these ROIs were then measured while subjects read moral stories from the current study. Within the ROI, the average percent signal change (PSC) relative to fixation (PSC = 100 × raw blood-oxygen-level-dependent (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 (Poldrack, 2006).

RESULTS

Theory of mind localizer experiment

A whole-brain random effects analysis of the data replicated results of previous studies using the same task (Saxe & Kanwisher, 2003), revealing a higher BOLD response during belief stories as compared to physical stories, in the RTPJ, LTPJ, dorsal (D), middle (M), and ventral (V) MPFC, and precuneus (PC) ($p < 0.05$, family-wise correction). ROIs were identified in individual subjects at the same threshold (Table 1): RTPJ (identified in 17 of 17 subjects), LTPJ (17/17), PC (17/17), DMPFC (14/17), MMPFC (11/17), and VMPFC (11/17).

Moral judgment: Behavioral results

Subjects evaluated the moral status of actions on a scale from no blame/praise (1) to a lot of blame/praise (4). Blame and praise judgments of harm and help scenarios, respectively, as well as reaction times (see "Supplementary information") were analyzed using separate 2 × 2 (outcome, negative/positive vs. neutral, by belief, negative/positive vs. neutral) repeated measures ANOVAs (Figure 3).

Harm

Predicted main effects of outcome and belief were observed. Agents producing negative outcomes were judged more morally blameworthy than those causing neutral outcomes, negative: 2.84, neutral: 2.15; $F(1, 11) = 26.9$, $p = 3.0 \times 10^{-4}$, partial $\eta^2 = .71$. Agents with "negative" beliefs were judged more morally blameworthy than those with "neutral" beliefs, negative: 3.28, neutral: 1.71; $F(1, 11) = 1.0 \times 10^2$, $p = 1.0 \times 10^{-6}$, partial $\eta^2 = .90$. There was no significant interaction between belief and outcome.

Judgments of negative outcomes were faster than of neutral outcomes, $F(1, 11) = 12.3$ $p = .005$, partial $\eta^2 = .53$; there was no effect of belief on reaction time. There was an interaction between belief and outcome, $F(1, 11) = 20.9$, $p = .001$, partial $\eta^2 = .66$, driven by a faster response to intentional harm (mean: 2.0 s, SD: 0.5) than the other conditions: accidental harm (mean: 2.3 s, SD: 0.6), attempted harm (mean: 2.5 s, SD: 0.6), or all-neutral (mean: 2.5 s, SD: 0.6).

Help

Predicted main effects of outcome and belief were observed. Agents producing positive outcomes were judged more morally praiseworthy than agents producing neutral outcomes, positive: 2.71, neutral: 2.20; $F(1, 11) = 42.9$, $p = 4.1 \times 10^{-5}$, partial $\eta^2 = .69$. Agents with "positive" beliefs were judged more morally praiseworthy than agents with "neutral" beliefs, positive: 2.98, neutral: 1.93; $F(1, 11) = 55.2$, $p = 1.3 \times 10^{-5}$, partial $\eta^2 = .77$. An interaction between outcome and belief was also observed, $F(1, 11) = 6.1$, $p = .03$, partial $\eta^2 = .36$, such that belief (neutral vs. positive) made a greater difference in the case of positive outcomes, as compared to neutral outcomes. That is, attempted help received little praise.

Judgments of positive beliefs (mean: 2.2 s) were faster than neutral beliefs, mean: 2.6 s; $F(1, 11) = 9.7$, $p = .01$, partial $\eta^2 = .47$; judgments of positive outcomes (mean: 2.2 s) were also faster than neutral outcomes, mean: 2.6 s; $F(1, 11) = 19.8$, $p = .001$, partial $\eta^2 = .64$. There was no interaction between belief and outcome.

Moral judgment: fMRI results

We calculated the average PSC from rest in each ROI for the critical segment of each story (22–26 s), at which point all the critical information (i.e., belief and outcome) for moral judgment was available (see "Supplementary information"). We expected the differential response to occur during this time, based on previous results, and the structure and timing of the stimuli (Young et al., 2007; Young & Saxe, 2008). As in the behavioral analyses, the neural responses for harm and help were analyzed using separate 2 × 2 outcome by belief repeated measures ANOVAs (Figure 4).

Harm

We replicated our previous results using similar stimuli (Young et al., 2007; Young & Saxe, 2008): a belief by outcome interaction in the RTPJ, $F(1, 16) = 6.6$, $p = .02$, partial $\eta^2 = .29$. Specifically, for negative

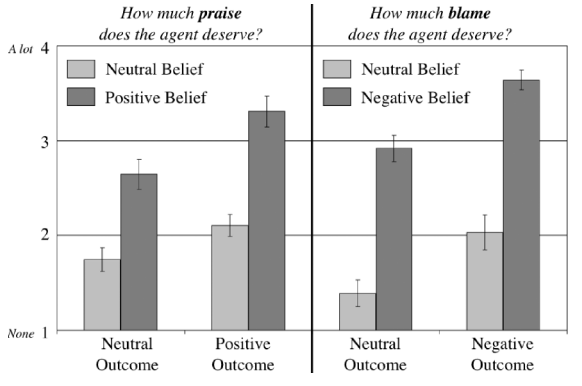

Figure 3. Moral praise and blame judgments. Error bars represent standard error.

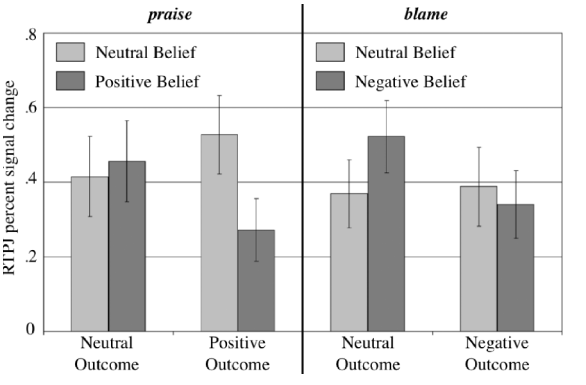

Figure 4. Percent signal change (PSC) from rest in the RTPJ for praise (left) and blame (right). Error bars represent standard error.

outcomes, there was no difference between neutral beliefs (mean PSC: 0.39) and negative beliefs (mean PSC: 0.34), but for neutral outcomes, there was a significant difference between neutral beliefs (mean PSC: 0.37) and negative beliefs, mean PSC: 0.52; $t(16) = 3.317$, $p = .004$. As in previous research (Young et al., 2007; Young & Saxe, 2008), planned comparisons also revealed that PSC for attempted harm was higher than for each of the other conditions, accidental harm: $t(16) = 2.6$, $p = .02$; intentional harm: $t(13) = –3.3$ $p = .004$).

Consistent with this ROI analysis, a random effects whole-brain analysis ($p < .001$, uncorrected) revealed greater activation of attempted harm (negative belief, neutral outcome) as compared to non-harm (neutral belief, neutral outcome) stories in the RTPJ (average peak voxel coordinates [56 –59 24]). No brain regions were found using a more stringent threshold, $p < .05$, family-wise correction, consistent with the higher power of functional ROI analyses to detect subtle but systematic response profiles (Saxe et al., 2006).)

A belief by outcome interaction was also observed in the LTPJ, $F(1, 16) = 17.5$, $p = .001$, partial $\eta^2 = .52$ (Figure 5); and DMPFC, $F(1, 16) = 5.7$, $p = .03$, partial $\eta^2 = .31$ (Figure 6). These effects were similar but less selective than those in the RTPJ: The LTPJ response showed differences between attempted harm and the true belief conditions, i.e., intentional harm, mean PSC: 0.35; $t(16) = –3.1$, $p = .007$, and all-neutral, mean PSC: 0.40; $t(16) = 3.9$, $p = .001$; but no difference between the two false belief conditions, i.e., attempted harm (mean PSC: 0.56) and accidental harm (mean PSC: 0.48; $p = .30$). In the DMPFC response, only a difference between attempted harm (mean PSC: 0.82) and intentional harm (mean PSC: 0.53) was observed, $t(16) = –2.4$, $p = .03$; the responses for all-neutral (mean PSC: 0.65) and accidental harm (mean PSC: 0.72) were intermediate. The PC, MMPFC, and VMPFC showed no significant effects.

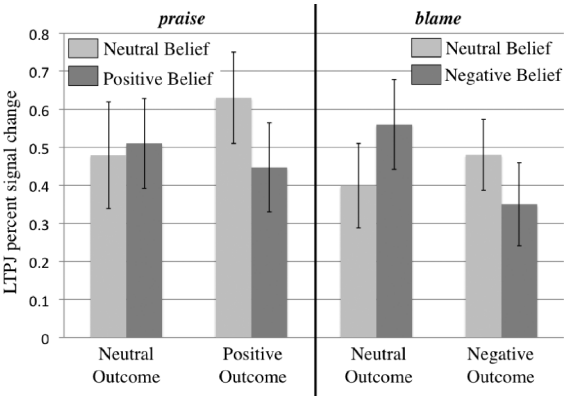

Figure 5. PSC from rest in the LTPJ for praise and blame. Error bars represent standard error.

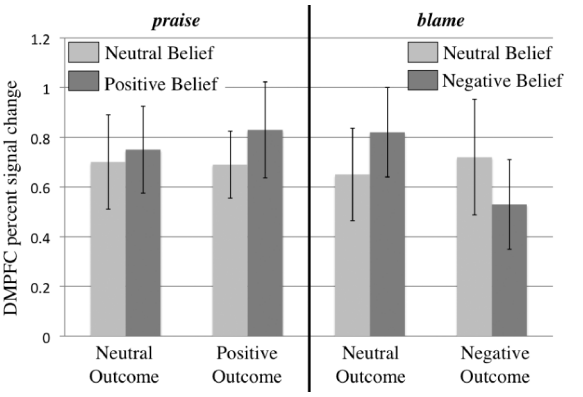

Figure 6. PSC from rest in the DMPFC for praise and blame. Error bars represent standard error.

Help

For the help cases, we observed a main effect of belief in the RTPJ, $F(1, 16) = 5.7$, $p = .03$, partial $\eta^2 = .26$. Importantly, this main effect was qualified by a belief by outcome interaction, complementary to the interaction observed in the harm cases, $F(1, 16) = 19.8$, $p = 4.0 \times 10^{-4}$, partial $\eta^2 = .55$: For positive outcomes, there was a difference between neutral beliefs, mean PSC: 0.53, and positive beliefs, mean PSC: 0.27; $t(16) = 4.1$, $p = .001$; for neutral outcomes, there was no difference between neutral beliefs (mean PSC: 0.41) and positive beliefs (mean PSC: 0.46). Consistent with this ROI analysis, a random effects whole-brain analysis ($p < .001$, uncorrected) revealed greater activation of accidental help (neutral belief, positive outcome) as compared to intentional help (positive belief, positive outcome) in the RTPJ (average peak voxel coordinates [60 –56 34]), though no brain regions were found using a more stringent threshold ($p < .05$, family-wise correction).

A belief by outcome interaction was also observed in the LTPJ, $F(1, 16) = 8.7$, $p = .009$, partial $\eta^2 = .35$: For neutral outcomes, there was no difference between neutral beliefs, mean PSC: 0.58, and positive beliefs, mean PSC: 0.51; and for positive outcomes, there was a difference between neutral beliefs, mean PSC: 0.63, and positive beliefs, mean PSC: 0.45; $t(16) = 2.4$, $p = .03$. A main effect of belief was observed in the MMPFC, $F(1, 16) = 5.9$, $p = .04$, partial $\eta^2 = .37$; a higher response was observed for positive beliefs (mean PSC: 0.38) than neutral beliefs (mean PSC: 0.21). The PC, DMPFC, and VMPFC showed no significant effects.

Harm vs. help

We found no main effect of harm vs. help in the RTPJ PSC, $t(16) = 0.4$, $p = .70$, or any other ROI except for the LTPJ. The LTPJ PSC was higher for help, mean PSC: 0.52, than for harm, mean PSC: 0.45, $t(16) = 2.8$, $p = .01$, but random effects whole-brain analyses of differential activation for positive (help) vs. negative (harm) scenarios yielded no significant clusters ($p < .05$, family-wise correction). Critically, though, the pattern of activation in the RTPJ was significantly different for help vs. harm scenarios with respect to the false belief cases (attempted help/harm and accidental help/harm). In particular, when we compared the difference for attempted vs. accidental help to attempted vs. accidental harm, the RTPJ response was greater for attempts than accidents in the case of harm and greater for accidents than attempts in the case of help, as predicted and as indicated by a significant interaction between condition (attempt vs. accident) and valence (harm vs. help) in a 2×2 repeated measures ANOVA, $F(1, 16) = 4.4$, $p = .05$, partial $\eta^2 = .21$.

GENERAL DISCUSSION

At the broadest level, the current study underscores the critical role for ToM in moral judgments of blame and praise. Both blame and praise judgments were influenced not only by the action's outcome but also the actor's mental state. Correspondingly, brain regions including the RTPJ, LTPJ, and MPFC, known to support ToM, were recruited across harm and help scenarios, indicating that blame and praise depend on computations occurring in the same neural substrates for ToM. The following discussion will therefore focus on how the neural evidence informs the specific role of mental states in morality, across blame and praise, and how the results specifically address the intuitive prosecutor vs. goal incompletion hypotheses. Is mental state information processed differently for different moral judgments, across harmful and helpful actions?

The role of the RTPJ in "intuitive prosecution"

The results of the current study replicate and resolve a previous puzzle about ToM in moral judgment. We aimed to test two alternative interpretations of a previously observed neural pattern: selective enhancement of RTPJ activation for failed attempts to harm (Young et al., 2007; Young & Saxe, 2008). On the intuitive prosecutor hypothesis, the enhanced activation reflects greater attention to or deeper processing of mental state information that supports a negative moral judgment. On the goal incompletion hypothesis, the enhanced activation reflects greater processing of salient goals that are not completed. Consistent with the intuitive prosecutor hypothesis, we found that the RTPJ response was greater for failed attempts than accidents in the case of harm, and greater for accidents than failed attempts in the case of help. More precisely, the RTPJ response discriminated between neutral and negative beliefs when the outcome was neutral (but not negative) for blame, and between neutral and positive beliefs when the outcome was positive (but not neutral) for praise. The RTPJ response may therefore reflect finer mental state discriminations when outcomes are neutral or positive, "working overtime" to detect "bad beliefs" especially when there's no other reason to blame or withhold praise from the agent. Participants thus served as "intuitive prosecutors" (Haidt, 2007; Tetlock, 2002), seeking mental state evidence to assign blame and withhold praise in morally ambiguous situations.

As such, these results are consistent with the broader phenomenon of moral rationalization: People search, post hoc, for evidence to support their moral judgments (Alicke, 2000; Gazzaniga, 2000; Haidt, 2001; Haidt, Koller, & Dias, 1993; Pizarro, Laney, Morris, & Loftus, 2006; Wheatley & Haidt, 2005). For example, when participants are unable to explain why they take incest to be morally wrong even in the absence of procreation or physical or emotional harm, they are "morally dumbfounded" (Haidt, 2001). At that point, participants often appeal to hypothetical harms or other invented consequences to rationalize their judgment.

An asymmetry between blame and praise

The current results show greater processing of mental states that support negative moral judgments, for assigning moral blame and withholding moral praise. These results relate specifically to other functional neuroimaging and behavioral research showing greater attention to mental states for negative vs. positive judgments. Prior behavioral work, for example, has shown that participants judge impulsive crimes (e.g., crimes of passion) as less morally blameworthy than deliberate or premeditated crimes but impulsive and deliberate charitable behavior as equally morally

praiseworthy (Pizarro et al., 2003). In other research, participants have been shown to attribute greater intent to agents bringing about negative vs. positive side-effects (Knobe, 2003, 2005). In one example, a CEO implements a profitable program, foreseeing that he will help/harm the environment as a side-effect of his action, though he has no intention to help/harm the environment. Participants judge the CEO as intentionally harming—but not helping—the environment.

Our own recent work has shown that participants appeal to mental state information especially when assigning moral blame (Kliemann, Young, Scholz, & Saxe, 2008; Young, Nichols, & Saxe, 2010c). When participants made *negative* moral judgments of *disliked* actors, they judged their harmful actions as more morally blameworthy and more intentional. These negative judgments were also accompanied by an increase in the RTPJ response, indicating greater processing of mental states for negative moral judgments. The neural evidence in the current study suggests that our participants engaged in more mental state reasoning when making negative moral judgments, assigning blame and withholding praise. Though we observed no overall effect of blame vs. praise, the detailed pattern of results suggests that neural substrates for processing mental states are recruited more robustly when mental states uniquely license negative moral judgments.

Reverse inference and other functions of the RTPJ

Our interpretation of the current results relies on a "reverse" inference, taking activity in a brain region (i.e., the RTPJ) to be evidence for the engagement of a specific cognitive process (i.e., extra mental state processing). The validity of a reverse inference depends on the prior evidence of the target brain region's selectivity for the cognitive process in question (Poldrack, 2006; Young & Saxe, 2009a). Of the regions implicated in ToM, the RTPJ appears to be especially selective for processing mental states such as beliefs, in and outside the moral domain (Perner et al., 2006; Saxe & Powell, 2006; Young & Saxe, 2008).

In other research, however, nearby regions have been implicated in attention to unexpected stimuli (Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000; Mitchell, 2007), including unexpected human actions (Buccino et al., 2007; Grezes, Frith, & Passingham, 2004; Pelphrey, Morris, & McCarthy, 2004), as well as other inconsistent information (Ferstl, Neumann, Bogler, & von Cramon, 2008; Simos, Basile, & Papanicolaou, 1997; Virtue, Parrish, & Jung-Beeman, 2008). Could the current results be due to differences in attention across the conditions (e.g., attempted/accidental harm/help) of the current study? We think this alternative unlikely for the following four reasons.

First, there is no a priori reason why attempted harm and accidental help (vs. accidental harm and attempted help, where mental state and outcome factors also conflict) should lead to more shifts of attention. All stimuli were presented verbally in similar language across conditions. Harm and help scenarios were also matched for word count across conditions. Moreover, shifts of attention are generally accompanied by slower reaction times, but we observed no reaction time differences between the critical conditions (e.g. Attempted Harm or Accidental Help, vs. All Neutral).

Second, a recent study, using higher resolution imaging and a bootstrap analysis, found a small but reliable separation between the peaks of functional regions for attention vs. ToM in higher resolution images (Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009), consistent with evidence from a recent meta-analysis (Decety & Lamm, 2007).

Third, in another recent fMRI study, participants read stories describing mental or physical states, which were unexpected or expected; the RTPJ response was significantly higher for mental vs. physical states but not sensitive to the difference between unexpected and expected stories in either domain (Young, Dodell-Feder, & Saxe, 2010b).

Finally, previously observed activations patterns for unexpected human actions have been mostly centered on the superior temporal sulcus (STS) rather than the functional region of the RTPJ for ToM; furthermore, processing unexpected (vs. expected) human actions may engage not only greater attention but greater ToM, that is, reasoning about the beliefs and intentions of the actor.

MPFC and social cognition

The DMPFC showed a similar but less selective pattern compared to the RTPJ for harm scenarios of the current task. Previous research suggests that the MPFC is recruited not for encoding explicit belief information (Saxe & Powell, 2006) but more broadly for moral cognition (Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007; Greene et al., 2004; Koenigs et al., 2007; Mendez, Anderson, & Shapira, 2005; Young et al., 2010) and social cognition (Adolphs, 2003; Mitchell, Macrae, & Banaji, 2006). Recent work suggests a role for the DMPFC in reasoning

about the desires or valenced attitudes of individuals dissimilar to oneself; by contrast, a more ventral region of MPFC was implicated in judging the desires/attitudes of individuals similar to oneself (Adolphs, 2003; Mitchell et al., 2006). It is therefore possible that the DMPFC activation reflects desire inferences—including negative desires in the case of attempted harm. The DMPFC did not, however, show selective effects in the case of positive desires, for the help scenarios. Future work should characterize the distinction between beliefs and desires, and the functional roles of different brain regions in processing different mental state content.

CONCLUSIONS

With few exceptions (Haidt, 2003; see Takahashi et al., 2008 for comparisons between emotional responses to "moral depravity" vs. "moral beauty"), prior cognitive neuroscience research has focused primarily on moral judgments of harmful actions, as well as other violations of moral norms (e.g., breaking promises, committing incest). The current study suggests that ToM processes may be disproportionately engaged when participants assign blame in the absence of a harmful outcome or withhold praise in the presence of a helpful outcome; that is, when participants become "intuitive prosecutors" and search for and attend to evidence in support of a (relatively) negative moral verdict.

In the future, the present paradigm may be useful for research into moral evaluations of ingroup vs. outgroup members. In the presence of a group boundary, participants may be differentially motivated to blame and praise and to take internal (mental state) vs. external information into account. Detailed understanding of the neural basis of moral blame and praise, ToM, and their relationship may then provide a window into complex social relations—both how they succeed and when they break down.

Manuscript received 7 May 2010
Manuscript accepted 14 September 2010
First published online 25 January 2011

REFERENCES

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

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126* (4), 556–574.

Aichhorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Do visual perspective tasks need theory of mind? *NeuroImage*, *30*(3), 1059–1068.

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*(5), 803–817.

Buccino, G., Baumgaertner, A., Colle, L., Buechel, C., Rizzolatti, G., & Binkofski, F. (2007). The neural basis for understanding non-intended actions. *NeuroImage*, *36*, Suppl. 2, T119–T127.

Ciaramelli, E., Muccioli, M., La Davas, 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*(13), 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*(3), 292–297.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition*, *108*(2), 353–380.

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*(12), 1082–1089.

de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254–1258.

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.

den Ouden, H. E., Frith, U., Frith, C., & Blakemore, S. J. (2005). Thinking about intentions. *Neuroimage*, *28*(4), 787–796.

Ferstl, E. C., Neumann, J., Bogler, C., & von Cramon, D. Y. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Hum Brain Mapp*, *29*(5), 581–593.

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*(2), 109–128.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *358*(1431), 459–473.

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*(1), 11–21.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, *123* (Pt 7), 1293–1326.

Gazzaniga, M. S. (2005). *The ethical brain*. New York, NY: Dana Press.

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*(11), 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.

Grezes, J., Frith, C. D., & Passingham, R. E. (2004). Inferring false beliefs from the actions of oneself and others: An fMRI study. *NeuroImage*, *21*(2), 744–750.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer & H. H. Goldsmith (Eds.), *Handbook of affective sciences.* (pp. 852–870). Oxford, UK: Oxford University Press.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*, 998–1002.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*(4), 613–628.

Harenski, C. L., & Hamaan, S. (2006). Neural correlates of regulating negative emotions related to moral violations. *NeuroImage*, *30*(1), 313–324.

Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. (2007). A dissociation between moral judgment and justification. *Mind and Language*, *22*, 1–21.

Heekeren, H. R., Wartenburger, I., Schmidt, H., Schwintowski, H. P., & Villringer, A. (2003). An fMRI study of simple ethical decision-making. *NeuroReport*, *14*, 1215–1219.

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science*, *320*(5879), 1092–1095.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion*, *9*(3), 435–439.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, *46*(12), 2949–2957.

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*, 309–324.

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., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*, 908–911.

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

Mikhail, J. M. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 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* (4), 655–663

Moll, J., de Oliveira-Souza, R., Moll, F. T., Ignacio, F. A., Bramati, I. E., Caparelli-Daquer, E. M., et al. (2005). The moral affiliations of disgust. *Journal of Cognitive Behavioral Neurology*, *18*(1), 68–78.

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), 15623–15628.

Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: The perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of Cognitive Neuroscience*, *16*(10), 1706–1716.

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience*, *1*(3–4), 245–258.

Pizarro, D. A., Laney, C., Morris, E. K., & Loftus, E. F. (2006). Ripple effects in memory: Judgments of moral blame can distort memory for events. *Memory and Cognition*, *34*(3), 550–555.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, *14*(3), 267–272.

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

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., & Kilts, C. (2002). A neural basis for social cooperation. *Neuron*, *35*(2), 395–405.

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*(11), 2475–2480.

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, *30*(4), 1088–1096; discussion 1097–1089.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*(4), 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*(8), 692–699.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*, *43*(10), 1391–1399.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One*, *4*(3), e4869.

Seligman, M. E., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, *55*(1), 5–14.

Simos, P. G., Basile, L. F., & Papanicolaou, A. C. (1997). Source localization of the N400 response in a sentence-reading paradigm using evoked magnetic fields and magnetic resonance imaging. *Brain Research*, *762*(1–2), 29–39.

Takahashi, H., Kato, M., Matsuura, M., Koeda, M., Yahata, N., Suhara, T., et al. (2008). Neural correlates of human virtue judgment. *Cerebral Cortex*, *18*(8), 1886–1891.
Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*(3), 451–471.
Virtue, S., Parrish, T., & Jung-Beeman, M. (2008). Inferences during story comprehension: Cortical recruitment affected by predictability of events and working memory capacity. *Journal of Cognitive Neuroscience*, *20*(12), 2274–2284.
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*(1), 170–181.
Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*(10), 780–784.
Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., & Damasio, A. (2010). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron*, *65*, 845–851.
Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010a). Disruption of the right temporo-parietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 6753–6758.
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 of the United States of America*, *104*(20), 8235–8240.
Young, L., Dodell-Feder, D., & Saxe, R. (2010b). What gets the attention of the temporo-parietal junction? An FMRI investigation of attention and theory of mind. *Neuropsychologia*, *48*, 2658–2664.
Young, L., Nichols, S., & Saxe, R. (2010c). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, *1*, 333–349.
Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*, 1912–1920.
Young, L., & Saxe, R. (2009a). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*(7), 1396–1405.
Young, L., & Saxe, R. (2009b). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072.

SUPPLEMENTARY INFORMATION

Word count

Word count was matched across harm conditions (mean ± *SD* for the all-neutral condition: 103 ± 10; accidental harm: 101 ± 9; attempted harm: 103 ± 10; intentional harm: 103 ± 9). On average, scenarios featuring negative beliefs contained the same number of words as scenarios featuring neutral beliefs, $F(1, 23) = 0.15$, $p = .70$, partial $\eta^2 = .006$; scenarios featuring negative outcomes contained the same number of words as scenarios featuring neutral outcomes scenarios, $F(1, 23) = 0.17$, $p = .68$, partial $\eta^2 = .007$.

Word count was also matched across help conditions, mean ± *SD* for the all-neutral condition: 94 ± 12; accidental help: 94 ± 12; attempted help: 94 ± 11; intentional help: 95 ± 11. On average, scenarios featuring positive beliefs contained the same number of words as scenarios featuring neutral beliefs, $F(1, 23) = 0.007$, $p = .94$, partial $\eta^2 < .001$; scenarios featuring positive outcomes contained the same number of words as scenarios featuring neutral outcome, $F(1, 23) = 0.32$, $p = .58$, partial $\eta^2 = .013$.

A 4 × 2 (condition, all-neutral vs. accident vs. attempt vs. intentional, by valence, harm vs. help) mixed effects ANOVA yielded no effect of condition, $F(2, 44) = 0.19$, $p = .83$, partial $\eta^2 = .008$, and no interaction between condition and valence, $F(2, 44) = 0.42$, $p = .66$, partial $\eta^2 = .02$, but a main effect between-subjects (NB: help and harm stories are different stories) of valence, $F(1, 45) = 6.6$, $p = .014$, partial $\eta^2 = .13$.

Reaction time

We found no difference between positive beliefs and negative beliefs, for either neutral outcomes, $t(11) = -0.39$, $p = .7$, or valenced (positive, negative) outcomes, $t(11) = 0.18$, $p = .9$. We also found no difference between positive outcomes and negative outcomes, for either neutral beliefs, $t(11) = 0.58$, $p = .6$, or valenced (positive, negative) beliefs, $t(11) = -0.18$, $p = .9$ (Figure 7).

FMRI analyses of the RTPJ responses in participants with/without behavioral data

To test whether the RTPJ results in the group of participants for whom we lost behavioral data ($N = 5$) differed from the results in the group of participants for whom we did not lose behavioral data ($N = 12$), we conducted two mixed-effects ANOVAs for help and harm including group as a between-subjects variable.

Harm

A 2 × 2 × 2 (outcome, negative vs. neutral, by belief, negative vs. neutral, by group, behavioral data lost vs. not lost) ANOVA revealed only the critical belief by outcome interaction, $F(1, 15) = 4.4$, $p = .05$, partial $\eta^2 = .23$. The group variable did not interact

with belief, $F(1, 15) = 0.3$, $p = .57$, partial $\eta^2 = .02$, outcome, $F(1, 15) = 0.08$, $p = .79$, partial $\eta^2 = .005$, or the belief by outcome interaction, $F(1, 15) = 0.2$, $p = .68$, partial $\eta^2 = .01$.

Help

Similarly, a 2 × 2 × 2 (outcome, positive vs. neutral, by belief, positive vs. neutral, by group, behavioral data lost vs. not lost) ANOVA revealed only the critical belief by outcome interaction, $F(1, 15) = 14.0$, $p = .002$, partial $\eta^2 = .48$. The group variable did not interact with belief, $F(1, 15) = 0.4$, $p = .54$, partial $\eta^2 = .03$, outcome, $F(1, 15) = 4.0$, $p = .06$, partial $\eta^2 = .2$, or the belief by outcome interaction, $F(1, 15) = 0.3$, $p = .58$, partial $\eta^2 = .02$.

Nevertheless, we conducted separate fMRI analyses for the 12 participants for whom we did not lose behavioral data, to ensure that the critical results obtained.

Harm

We observed a belief by outcome interaction in the RTPJ, $F(1, 11) = 4.8$, $p = .05$, partial $\eta^2 = .30$. Specifically, for negative outcomes, there was no difference between neutral beliefs, mean PSC: 0.46, and negative beliefs, mean PSC: 0.39; $t(11) = -1.02$, $p = .33$, but for neutral outcomes, there was a significant difference between neutral beliefs, mean PSC: 0.44, and negative beliefs, mean PSC: 0.59; $t(11) = 2.43$, $p = .03$. Planned comparisons also revealed that PSC for attempted harm was higher than for each of the other conditions, accidental harm: $t(11) = 1.76$, $p = .05$, one-tailed; intentional harm: $t(11) = -3.4$, $p = .005$.

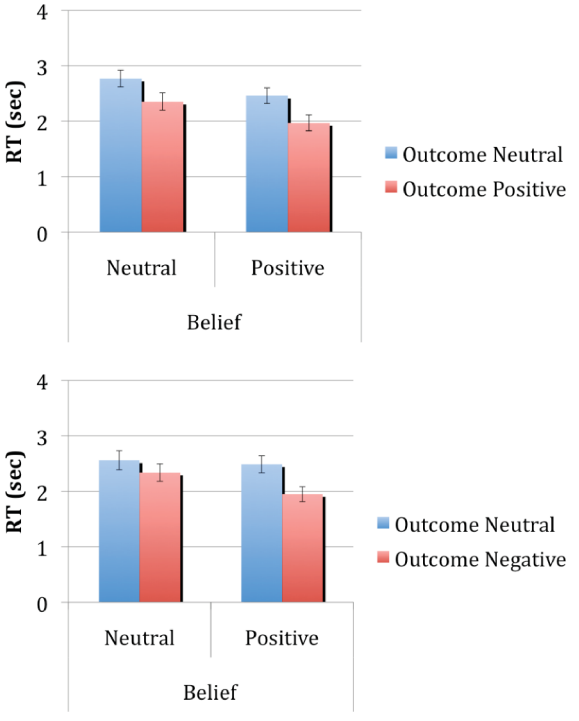

Figure 7. Reaction time data.

Help

For the help cases, we observed the critical belief by outcome interaction, $F(1, 11) = 30.7$, $p < .001$, partial $\eta^2 = .74$. For positive outcomes, there was a difference between neutral beliefs, mean PSC: 0.56, and positive beliefs, mean PSC: 0.31; $t(11) = -3.5$, $p = .005$; for neutral outcomes, there was no difference between neutral beliefs, mean PSC: 0.50, and positive beliefs, mean PSC: 0.57; $t(11) = 1.38$, $p = .20$.

Harm vs. Help

We found no main effect of harm vs. help in the RTPJ PSC, $t(11) = 0.27$, $p = .79$. The interaction between condition (attempt vs. accident) and valence (harm vs. help) in a 2 × 2 repeated measures ANOVA did not reach significance, $F(1, 11) = 0.933$, $p = .35$, partial $\eta^2 = .08$.

Non-significant main effects in the RTPJ

Harm

A 2 × 2 (outcome, negative vs. neutral, by belief, negative vs. neutral) ANOVA yielded a nonsignificant main effect of belief, $F(1, 16) = 0.24$, $p = .14$, partial $\eta^2 = .12$, and outcome, $F(1, 16) = 4.0$, $p = .06$, partial $\eta^2 = .20$.

Help

A 2 × 2 (outcome, positive vs. neutral, by belief, positive vs. neutral) ANOVA yielded a nonsignificant main effect of outcome, $F(1, 16) = 0.35$, $p = .56$, partial $\eta^2 = .02$.

FMRI analyses of the RTPJ response over third and fourth segment

To analyze the RTPJ results over the third segment (i.e., when the action and outcome were made available) and

fourth segment (i.e., when a moral judgment was elicited), we calculated the average PSC from rest in each region of interest ROI for these segments of each story.

Harm

A 2 × 2 (outcome, negative vs. neutral, by belief, negative vs. neutral) ANOVA revealed a main effect of outcome, $F(1, 16) = 5.8$, $p = .03$, partial $\eta^2 = .27$, and a marginal belief by outcome interaction, $F(1, 16) = 2.6$, $p = .12$, partial $\eta^2 = .14$. The PSC for attempted harm, mean PSC: 0.27, was higher than intentional harm, mean PSC: 0.13; $t(16) = -3.8$, $p = .002$, though not significantly higher than the other conditions: accidental harm, mean PSC: 0.19, all-neutral, mean PSC: 0.22.

Help

A 2 × 2 (outcome, negative vs. neutral, by belief, negative vs. neutral) ANOVA revealed a main effect of belief, $F(1, 16) = 5.3$, $p = .04$, partial $\eta^2 = .25$, and a marginal belief by outcome interaction, $F(1, 16) = 8.8$, $p = .009$, partial $\eta^2 = .35$. The PSC for accidental help, mean PSC: 0.27, was higher than for intentional help, mean PSC: 0.09; $t(16) = -3.1$, $p = .007$, though not significantly higher than the other conditions: attempted help, mean PSC: 0.21; all-neutral, mean PSC: 0.22.