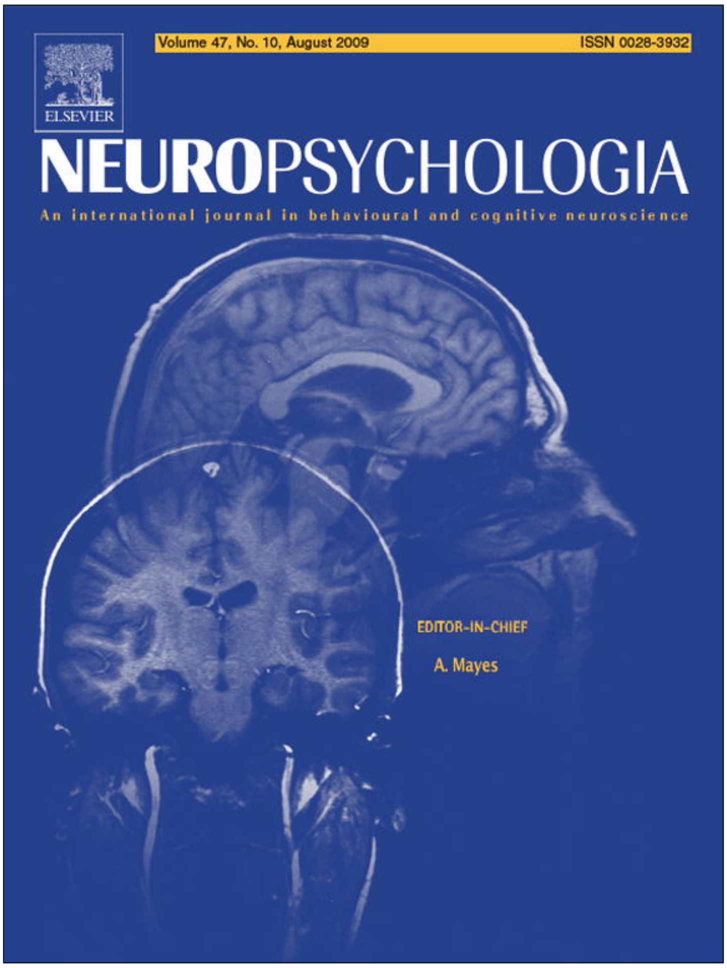
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Neuropsychologia 47 (2009) 2065–2072

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Innocent intentions: A correlation between forgiveness for accidental harm and neural activity☆

Liane Young*, Rebecca Saxe

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 2 October 2008
Received in revised form 12 March 2009
Accepted 24 March 2009
Available online 5 April 2009

Keywords:
Morality
Theory of mind
Belief attribution
Exculpation
Forgiveness
fMRI
Temporo-parietal junction
Ventromedial prefrontal cortex

ABSTRACT

Contemporary moral psychology often emphasizes the universality of moral judgments. Across age, gender, religion and ethnicity, people's judgments on classic dilemmas are sensitive to the same moral principles. In many cases, moral judgments depend not only on the outcome of the action, but on the agent's beliefs and intentions at the time of action. For example, we blame agents who attempt but fail to harm others, while generally forgiving agents who harm others accidentally and unknowingly. Nevertheless, as we report here, there are individual differences in the extent to which observers exculpate agents for accidental harms. Furthermore, we find that the extent to which innocent intentions are taken to mitigate blame for accidental harms is correlated with activation in a specific brain region during moral judgment. This brain region, the right temporo-parietal junction, has been previously implicated in reasoning about other people's thoughts, beliefs, and intentions in moral and non-moral contexts.

© 2009 Elsevier Ltd. All rights reserved.

Father, forgive them, for they know not what they do. Luke 23:34

1. Introduction

Classic moral dilemmas often require an observer to judge whether it is permissible to harm one innocent person to save many. For example, is it permissible to push a man off a bridge so that his body will stop a trolley from running over five other people? Competition between emotional aversion to committing harm (e.g., pushing the man), and abstract reasoning, in this case, utilitarian reasoning about maximizing aggregate welfare (e.g., five lives are worth more than one), gives rise to the 'dilemma', and to characteristic neural response profiles (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004). These results have led to two-process theories of moral judgment (Cushman, Young, & Hauser, 2006; Greene et al., 2004; Haidt, 2001; Hsu, Anen, & Quartz, 2008). Implicit, automatic processes lead observers to reject emotionally aversive harms. Explicit, controlled processes support abstract reasoning and cognitive control.

Here, we extend two-process theories by considering a third factor upon which many moral judgments depend: the agent's mental state. When we evaluate an action, be it killing one or letting many die, harming or helping, breaking the law, breaking a promise, or breaking fast with the wrong sorts of people, we consider the agent's mental state at the time of her action. Did she know what she was doing? Did she act intentionally or accidentally? Observers judge intentional harms as worse than accidental harms (e.g., Cushman, 2008). Observers are even sensitive to more subtle mental state distinctions, judging harms intended as necessary means to an end to be worse than harms that are merely foreseen as side-effects of one's action (Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006; Cushman et al., 2006; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2007).

Observers differ in the degree to which they take mental states into account for moral judgments. For example, children 5 years old and younger rely primarily on the action's observable outcomes (Hebble, 1971; Piaget, 1965/1932; Shultz, Wright, & Schleifer, 1986; Yuill, 1984; Yuill & Perner, 1988; Zelazo, Helwig, & Lau, 1996). Children are particularly unlikely to mitigate blame for accidental harms, and even judge accidental harms to be worse than failed attempts to harm (e.g., Baird & Astington, 2004). Not until they are 6 or 7 years old do children begin to make moral judgments that depend substantially on beliefs (Baird & Astington, 2004; Baird & Moses, 2001; Darley & Zanna, 1982; Fincham & Jaspers, 1979;

☆ This study was carried out at the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

* Corresponding author at: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Building 46, Room 4021, 43 Vassar Street, Cambridge, MA 02139, USA. Tel.: +1 617 312 5544; fax: +1 617 324 2890.
E-mail address: lyoung@mit.edu (L. Young).

0028-3932/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2009.03.020

Author's personal copy

Karniol, 1978; Shultz et al., 1986; Yuill, 1984) and integrate the distinct outcome and mental state features of actions (Grueneich, 1982; Weiner, 1995; Zelazo et al., 1996). There is also evidence that even adult observers differ in the extent to which they exculpate an agent for accidentally causing harm, and the extent to which they appeal to mental state factors in doing so (e.g., Cohen & Rozin, 2001; Nichols & Ulatowski, 2007).

In the current study, we investigated the neural correlates of individual differences in moral judgments that depend on agents' beliefs about whether or not they will cause harm. Consider a case in which an agent mistakes some poisonous white substance for sugar and, as a result, accidentally makes her friend sick by putting the poisonous substance in her coffee. Here, the agent believes falsely that her action will be harmless, and it is her false belief leads her to cause harm in spite of innocent intentions. Nevertheless, observers may disagree about the amount of blame that she deserves. Young children, and even some adults, may consider the agent very morally blameworthy for making her friend sick, in spite of her innocent intentions.

The neural mechanisms for reasoning about beliefs (or, more generally, mental states) have been investigated in a series of recent functional magnetic resonance imaging (fMRI) studies. These studies reveal a consistent group of brain regions for mental state reasoning in non-moral contexts: the medial prefrontal cortex, right and left temporo-parietal junction, and precuneus (Ciaramidaro et al., 2007; Fletcher et al., 1995; Gallagher et al., 2000; Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007; Ruby & Decety, 2003; Saxe & Kanwisher, 2003; Vogeley et al., 2001). Of these regions, the right temporo-parietal junction (RTPJ) in particular appears to be selective for belief attribution (Aichorn, Perner, Kronbichler, Staffen, & Ladurner, 2005; Fletcher et al., 1995; Gallagher et al., 2000; Gobbini et al., 2007; Perner, Aichorn, Khronblicher, Staffen, & Ladurner, 2006; Saxe & Wexler, 2005). For example, the response in the RTPJ is high when subjects read stories about a character's thoughts, beliefs, knowledge but low during stories containing other socially relevant information, for example, a character's physical or cultural traits, or even internal sensations such as hunger (Saxe & Powell, 2006).

Recently, we have also investigated the neural basis of belief reasoning in moral contexts (Young, Cushman, Hauser, & Saxe, 2007; Young & Saxe, 2008; Young & Saxe, in press). While in the scanner, participants read stories about a protagonist, and made moral judgments about the protagonist's actions. During the story, participants read two kinds of morally relevant information: (1) the protagonist's belief (e.g., that the powder was sugar) and (2) the reality (e.g., that the powder was poison). We investigated the neural response while participants initially processed these pieces of information. We found that the response in the RTPJ and precuneus was higher while participants read about beliefs than about other facts, independent of the order in which belief and non-belief facts were presented (Young & Saxe, 2008). However, this initial encoding response did not distinguish between negative and neutral beliefs (e.g., that the powder was poison versus sugar), between true and false beliefs, or between negative and neutral outcomes. In the current paper, we investigated a different question: namely, which brain region's response predicts people's use of belief information during the moral judgment itself?

We predicted that participants' use of belief information to make moral judgments would be correlated with the recruitment of specific brain regions associated with mental state reasoning. More specifically, we predicted that higher activation in these brain regions would lead to less blame (or more exculpation) for accidental harm, and more blame for attempted harm. Given prior evidence for its selectivity, we specifically predicted that these patterns would be observed in the RTPJ.

2. Methods

Fifteen right-handed neurologically normally adults (aged 18–22 years, 8 women, 7 men) participated in the study for payment. All participants were native English speakers, had normal or corrected-to-normal vision, and gave written informed consent in accordance with the requirements of Internal Review Board at MIT. Participants were scanned at 3T (at the MIT scanning facility in Cambridge, MA) using twenty-six 4-mm-thick near-axial slices covering the whole brain. Standard echoplanar imaging procedures were used (TR = 2 s, TE = 40 ms, flip angle 90°).

The experiment followed a 2×2 design. Stimuli consisted of 4 variations (conditions) of 24 moral scenarios (Fig. 1, see Supplementary Material for full text of all scenarios):

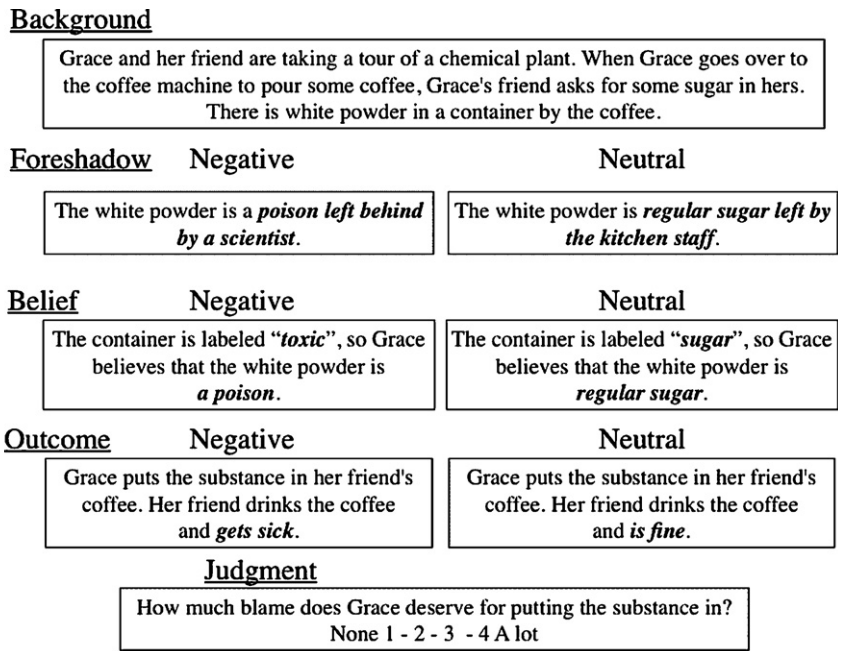


Fig. 1. Experimental stimuli and design. "Foreshadow" information foreshadows whether the action will result in a negative or neutral outcome. "Belief" information states whether the protagonist holds a belief that she is in a negative situation and that action will result in a negative outcome ("negative" 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. "Judgment" was presented alone on the screen for 4 s.

Author's personal copy

(i) Protagonists either harmed another person (negative outcome) or did no harm (neutral outcome).
(ii) Protagonists either believed that they were causing harm ("negative" belief) or believed they were causing no harm ("neutral" belief).

Each possible belief was true for one outcome and false for the other outcome; the agent held true beliefs in the no harm and intentional harm conditions and false beliefs in the accidental harm and attempted harm conditions. Word count was matched across conditions (mean ± S.D. 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 ($F(1, 23) = 0.15$ $p = 0.70$, partial $h^2 = 0.006$); scenarios featuring negative outcomes contained the same number of words as scenarios featuring neutral outcomes ($F(1, 23) = 0.17$ $p = 0.68$, partial $h^2 = 0.007$).

Stories were presented in four cumulative segments (previous segments remained on the screen when later segments were added): (1) background information to set the scene (0–6 s), (2) facts foreshadowing the eventual outcome (6–12 s), (3) the protagonist's belief (12–18 s), (4) the protagonist's action and its outcome (18–26 s). All of the story text was then removed from the screen, and replaced with the question and response scale. Subjects had 4 s (while the question was on the screen) to judge how much moral blame the protagonist deserved for performing a particular action on a 4-point scale (1: none, 4: a lot), using a button press. Subjects saw one version of each scenario. Stories were presented in a pseudorandom order; conditions were counterbalanced across runs and subjects. Fixation blocks (14 s) were interleaved between stories.

In the same scan session, subjects participated in four runs of a theory of mind (mental state reasoning) localizer experiment, contrasting stories requiring inferences about mental states (e.g., thoughts, beliefs) versus physical representations (e.g., outdated photographs, maps, signs; Saxe & Kanwisher, 2003).

3. FMRI analysis

MRI data were analyzed using SPM2 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. Each subject's data were motion corrected and 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, the event onset defined by the onset of text on screen.

Both whole-brain and tailored regions of interest (ROI) analyses were conducted. Six ROIs were defined for each subject individually based on a whole brain analysis of the independent localizer experiment, and defined as contiguous voxels that were significantly more active ($p < 0.001$, uncorrected, $k > 20$) while the subject read the mental state stories, as compared with the physical representation stories. All peak voxels are reported in MNI coordinates.

The responses of these ROIs were then measured while subjects read the moral stories from the current study. Within the ROI, the average percent signal change (PSC) relative to rest baseline (PSC = 100 × raw BOLD magnitude for (condition − fixation)/raw BOLD magnitude for fixation) was calculated for each condition at each time point (averaging across all voxels in the ROI and all blocks of the same condition). We then averaged together the time points within the judgment phase (30–34 s after story onset, to account for hemodynamic lag) to get a single PSC value for each region in each subject (Poldrack, 2006). This value was used in all analyses reported below.

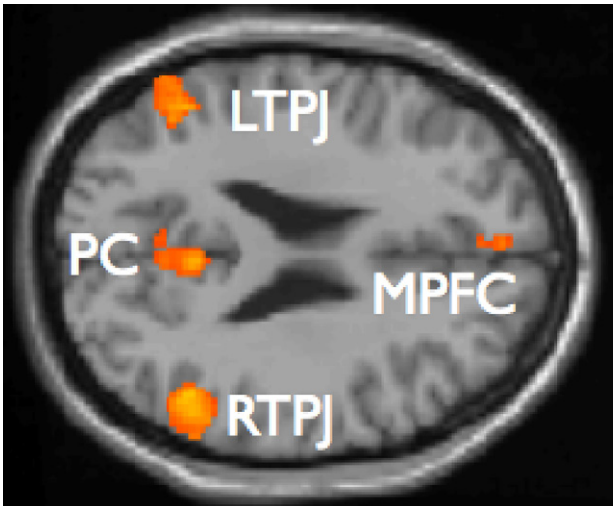


Fig. 2. Functional localizer results. Brain regions where the BOLD signal was higher for (nonmoral) stories about mental states than (nonmoral) stories about physical representations ($N = 15$, random effects analysis, $p < 0.001$, uncorrected, $k > 20$). These data were used to define regions of interest (ROIs).

Table 1
Localizer experiment results.

ROI	Individual ROIs			Whole-brain contrast		
	x	*y*	*z*	*x*	*y*	*z*
RTPJ	58	−56	23	56	−52	22
PC	1	−58	41	0	−54	40
LTPJ	−53	−58	26	−58	−58	24
dMPFC	1	56	38	4	56	40
mMPFC	1	58	17	−2	52	22
vMPFC	1	58	−14	−14	54	−12

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

4. Results

4.1. 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 the mental state as compared to physical representation stories, in the RTPJ, LTPJ, dorsal (D), middle (M), and ventral (V) MPFC, and precuneus (PC) ($p < 0.001$, uncorrected, $k > 20$). These regions of interest (ROIs) were identified in individual subjects at the same threshold (Fig. 2, Table 1): RTPJ (identified in 15 of 15 subjects), LTPJ (15/15), PC (15/15), DMPFC (13/15), MMPFC (10/15), and VMPFC (10/15).

4.2. Moral judgment: behavioral results

4.2.1. Moral judgment

Moral judgment data were analyzed in a 2 × 2 repeated measures ANOVA (Belief: neutral versus negative × Outcome: neutral

Table 2
Behavioral results.

	All-neutral	Accidental harm	Attempted harm	Intentional harm
Moral judgment (mean, S.D.)	1.5 (0.6)	2.1 (0.7)	2.9 (0.5)	3.7 (0.3)
Reaction time (mean, S.D.)	2.6 (0.5)	2.4 (0.4)	2.6 (0.5)	2.1 (0.4)

Average moral judgments and reaction times (mean, standard deviation) for each of the four experimental conditions. Moral judgments were given by subjects on a four-point scale (1, no blame, 4, a lot of blame). Reaction time was measured in seconds.

Author's personal copy

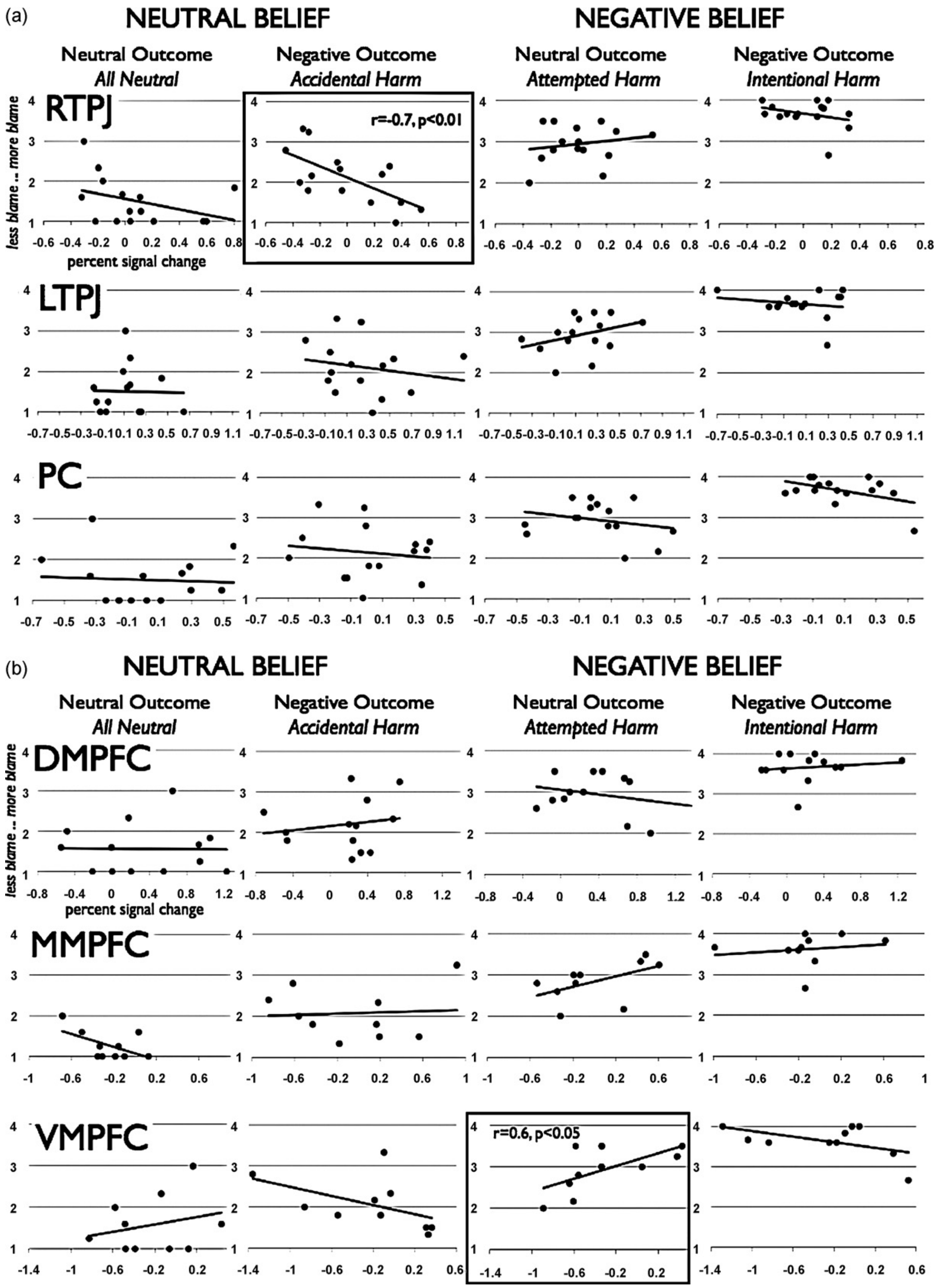


Fig. 3. (a) Individual differences in average percent signal change (PSC) in the RTPJ, LTPJ, and PC during moral judgment and average moral judgment (blame) for all four conditions (all-neutral, accidental harm, attempted harm, intentional harm). (b) Individual differences in average percent signal change (PSC) in the dorsal, middle, and ventral MPFC during moral judgment and average moral judgment (blame) for all four conditions (all-neutral, accidental harm, attempted harm, intentional harm).

Author's personal copy

Table 3
Correlations between average percent signal (PSC) in each ROI in each condition.

	All-neutral	Accidental harm	Attempted harm	Intentional harm
RTPJ	$r(15) = -0.38$, $p = 0.17$	$\mathbf{r(15) = -0.66}$, $\mathbf{p = 0.007}$	$r(15) = 0.19$, $p = 0.50$	$r(15) = -0.30$, $p = 0.28$
PC	$r(15) = -0.06$, $p = 0.82$	$r(15) = -0.14$, $p = 0.62$	$r(15) = -0.24$, $p = 0.38$	$r(15) = -0.46$, $p = 0.08$
LTPJ	$r(15) = -0.02$, $p = 0.94$	$r(15) = -0.19$, $p = 0.49$	$r(15) = 0.35$, $p = 0.20$	$r(15) = -0.18$, $p = 0.52$
DMPFC	$r(13) = -0.02$, $p = 0.96$	$r(13) = 0.18$, $p = 0.56$	$r(13) = -0.28$, $p = 0.36$	$r(13) = -0.11$, $p = 0.65$
MMPFC	$r(10) = -0.55$, $p = 0.10$	$r(10) = 0.07$, $p = 0.85$	$r(10) = 0.52$, $p = 0.12$	$r(10) = 0.17$, $p = 0.65$
VMPFC	$r(10) = 0.26$, $p = 0.47$	$r(10) = -0.49$, $p = 0.15$	$\mathbf{r(10) = 0.64}$, $\mathbf{p = 0.048}$	$r(10) = -0.51$, $p = 0.13$

Significant ($p < 0.05$) values bolded.

versus negative; Table 2). Protagonists causing harm were judged more blameworthy than those causing no harm ($F(1, 14) = 38.3$ $p = 2.4 \times 10^{-5}$, partial $h^2 = 0.73$). Protagonists with "negative" beliefs were judged more blameworthy than those with "neutral" beliefs ($F(1, 14) = 1.2 \times 10^2$ $p = 3.9 \times 10^{-8}$, partial $h^2 = 0.90$). We observed no belief by outcome interaction ($F(1, 14) = 0.37$ $p = 0.55$, partial $h^2 = 0.03$).

There was no effect of gender on moral judgment, when gender was included in the analysis. Gender did not interact with either belief ($F(1, 13) = 0.12$ $p = 0.73$, partial $h^2 = 0.01$) or outcome ($F(1, 13) = 0.03$ $p = 0.87$, partial $h^2 = 0.002$) or the interaction of belief and outcome ($F(1, 13) = 0.37$ $p = 0.53$, partial $h^2 = 0.03$).

4.2.2. Reaction time

Reaction time data were analyzed in the same fashion as the moral judgment data (Table 2). On average, judgments of protagonists causing negative outcomes were faster than judgments of protagonists causing neutral outcomes ($F(1, 14) = 16.2$ $p = 0.001$, partial $h^2 = 0.54$). There was no difference in reaction time for neutral and negative beliefs ($F(1, 14) = 3.0$ $p = 0.11$, partial $h^2 = 0.18$). We observed an interaction between belief and outcome ($F(1, 14) = 5.4$ $p = 0.04$ partial $h^2 = 0.28$); specifically, reaction times were longer for neutral beliefs than for negative beliefs for negative outcomes, but no different for neutral and negative beliefs for neutral outcomes.

There was no effect of gender on reaction time, when gender was included in the analysis. Gender did not interact with belief ($F(1, 13) = 0.55$ $p = 0.47$, partial $h^2 = 0.04$) or outcome ($F(1, 13) = 0.002$ $p = .97$, partial $h^2 < 0.001$) or the interaction between belief and outcome ($F(1, 13) = 0.35$ $p = 0.57$, partial $h^2 = 0.03$).

4.3. Moral judgment: fMRI results

For each of the four conditions, in each ROI, we calculated the correlation across participants between the average percent signal change (PSC) during the moral judgment (see analyses in Supplementary Material, Supplementary Fig. 1) and the value of the moral judgments (Fig. 3a and b).

The average PSC in the RTPJ during judgment of accidental harms was strongly and negatively correlated with the participant's average judgment of those accidental harms ($r = -0.66$, $p = 0.007$; Fig. 3a, top panel). That is, participants with high RTPJ activation assigned less blame, and participants with low RTPJ activation assigned more blame, to the same agents for the same actions resulting in the same harmful outcomes. The correlation effect was specific to the condition of accidental harm. There was no correlation between RTPJ response and judgments in the other conditions (all-neutral: $r = -0.38$, $p = 0.17$; attempted harm: $r = 0.19$, $p = 0.50$; intentional harm: $r = -0.30$, $p = 0.28$). There was also no correlation between RTPJ response and reaction time ($r = -0.32$, $p = 0.25$).

The correlation between moral judgment of accidental harms and PSC was specific to the RTPJ and did not emerge for any other ROI (Table 3). In the LTPJ (Fig. 3a, middle panel), there was no significant correlation between PSC and moral judgments in any condition (all $|r| < 0.40$; all $p > 0.20$). In the PC (Fig. 3a, bottom panel), there was no significant correlation between PSC and moral judgments in any condition (all $|r| < 0.50$; all $p > 0.05$). In the DMPFC (Fig. 3b, top panel), there was no significant correlation between PSC and moral judgments in any condition (all $|r| < 0.30$; all $p > 0.30$). In the MMPFC (Fig. 3b, middle panel), there was no significant correlation between PSC and moral judgments in any condition (all $|r| < 0.60$; all $p > 0.10$).

In the VMPFC (Fig. 3b, bottom panel), there was no correlation between PSC and moral judgments in the accidental harm ($r = -0.49$, $p = 0.15$), all-neutral ($r = 0.26$, $p = 0.47$), and intentional harm ($r = -0.51$, $p = 0.13$) conditions. There was, however, a weak positive correlation between PSC in the VMPFC and moral judgment of attempted harms ($r = 0.64$, $p = 0.048$). This correlation, however, would not survive a correction for multiple comparisons. There was no correlation between VMPFC response and reaction times ($r = -0.24$, $p = 0.50$) in this condition.

We also tested the significance of the difference between the correlation between moral judgment of accidental harms and response in the RTPJ and the same correlation in the other ROIs, from the same samples (e.g., to test for the difference between the RTPJ and the DMPFC, we included only those subjects in whom we identified both RTPJ and DMPFC; Chen & Popovich, 2002). We found significant differences between the RTPJ and the PC ($t = 3.65$, $p < 0.005$), LTPJ ($t = 2.23$, $p < 0.025$), and the DMPFC ($t = 2.86$, $p < 0.01$). The differences did not reach significance for the MMPFC ($t = 1.78$, $0.05 < p < 0.10$) or the VMPFC ($t = 1.63$, $0.05 < p < 0.10$).

We also conducted whole brain random effects analyses of the moral judgment experiment, to identify voxels in which the BOLD response at the time of the judgment was significantly correlated with the participant's moral judgment ($p < 0.001$, uncorrected, $k > 20$), for accidental or attempted harm. No brain regions showed a significant correlation with moral judgments in these analyses. These results are consistent with the higher power of functional ROI analyses to detect subtle but systematic response profiles (Saxe, Brett, & Kanwisher, 2006).

5. Discussion

FMRI findings have indicated that specific brain regions, including especially the RTPJ, support the ability to attribute beliefs to agents in both moral (e.g., Young et al., 2007; Young & Saxe, 2008; Young & Saxe, in press) and non-moral contexts (e.g., Saxe & Kanwisher, 2003; Perner et al., 2006). Behavioral studies have revealed that moral judgments depend significantly on mental state attribution; judgments of *moral blame* in particular depend on both the mental state (e.g., belief) of the agent and the outcome of the action (Cushman, 2008). The current results integrate these findings, showing that as the response in the RTPJ increases, so does the influence of belief information on moral judgments. More specifically, the extent to which the RTPJ is recruited during moral judgment is variable across subjects, and individual differences in the RTPJ response are correlated with the extent to which subjects use belief information in moral exculpation: subjects with higher activation of the RTPJ are more likely to exculpate agents for causing harm accidentally on the basis of a false belief. The current behav-

Author's personal copy

ioral and neural results therefore reinforce and clarify the role of mental state reasoning in moral judgment. Exculpating an agent who causes harm accidentally – an especially difficult task for young children – requires an especially robust mental state representation.

6. Moral universals and individual differences

Contemporary moral psychology often emphasizes the robustness of moral judgments to cultural and demographic differences: people are sensitive to the same moral principles independent of gender, age, ethnicity, and religion (e.g., O'Neill & Petrinovich, 1998; Petrinovich, O'Neill, & Jorgensen, 1993; Hauser et al., 2007). For example, the majority of subjects across cultures and demographic groups judge that it is permissible to turn a trolley away from five people and onto one person instead but impermissible to push a man off a footbridge so that his body stops a trolley from hitting five other people.

Nevertheless, there is evidence for systematic individual differences in moral judgment of these classic dilemmas. For example, the extent to which subjects engage in "cognitive" versus intuitive/emotional processes may influence judgments on "the trolley problem" (Greene et al., 2001, 2004). Individuals with higher working memory capacity are more likely to endorse utilitarian moral choices (Moore, Clark, & Kane, 2008), harming a few to save many. Individuals who score high in "need for cognition" (Cacioppo, Petty, & Kao, 1984) and low on "faith in intuition" (Epstein, Pacini, DenesRaj, & Heier, 1996) show the same pattern. Other research has shown that individual differences in "need for cognition" are related to individual differences in punitive attitudes, especially towards negligent or reckless behavior, such as drunk driving (Sargent, 2004). More specifically, individuals with a low "need for cognition" punish such behaviors more harshly, perhaps responding automatically to emotionally salient harmful outcomes; individuals with a high "need for cognition" respond more leniently, perhaps taking into account other factors such as situational or mental state factors and their interaction.

There is also some prior evidence of differences in the weight given specifically to mental states in moral judgment. One example is attitudes towards duty and obligation, when the agent's mental state is at odds with the agent's action. In one study of cultural differences (Cohen & Rozin, 2001), Jews were more likely to make outcome-based judgments, recognizing, for example, the value of taking care of one's parents even in the absence of positive feelings towards them. Christians, in contrast, were more likely to make mental state-based judgments, judging that caring for one's parents without appropriately positive mental states is hypocritical.

Related research has found individual differences in the *side-effect effect* or the Knobe effect (for a review, see Knobe, 2005). Participants are told that the chairman of a company implements a program to gain profit, but a side effect of that action, which he foresees but "doesn't care" about, is that the environment will be harmed. On average, participants judge that the chairman intentionally harmed the environment. By contrast, if the side-effect is that the environment will be *helped*, participants judge that he did not intentionally help the environment. The correct interpretation of the side-effect effect remains controversial (e.g., Machery, 2008). For our purposes, though, the key point is that there are individual differences in such judgments of intentionality and moral blame; some participants focus mostly on the chairman's desires, while others focus on what he *believed or knew* would happen (Nichols & Ulatowski, 2007).

In sum, a growing body of research suggests that there are individual differences in moral judgments, generally, and moral judgments based on beliefs, specifically. The current research fits nicely with this trend. Adults in our study differed in the extent to which they exculpated someone for an accident based on a false belief. Important questions for future research include whether these individual differences are stable across stimuli and experiments, and whether they extend to all domains of morality or are restricted to cases of bodily harm.

7. Accidents versus attempts

One open question concerning these results is: why was the response in the RTPJ correlated with the use of beliefs for moral judgments of accidental harms but not attempted harms? One possibility is that we simply had less power to detect the correlation in the attempted harm condition, because there was less variance across participants in moral judgments of attempted harms. An alternative, however, is that there are meaningful differences in the cognitive processes involved in using belief information to decrease, versus increase, moral blame. There are at least three ways in which belief information might be used differently in these two conditions: accidents and attempts.

First, moral judgments of accidents and attempts may depend on qualitatively different mental state attributions. Pilot behavioral data suggest that judgments of accidental harms depend mostly on what the protagonist *thought* or *didn't know*; by contrast, judgments of attempted harms depend on what the protagonist *desired* or *intended* (e.g., if the agent *believes* the stuff to be poison and puts it in her friend's coffee, she most likely *wants* to poison her friend). This distinction is consistent with the unexpected correlation we observed between judgments of attempted harms, and the BOLD response in the VMPFC. Prior neuroimaging and neuropsychological research studies (Greene et al., 2001; Koenigs et al., 2007; Mendez, Anderson, & Shapira, 2005; Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007) suggest that the VMPFC supports the processing of social and moral emotions for moral judgment. For example, individuals with lesions to the VMPFC are less sensitive to differences in the emotional salience and intentional nature of actions, choosing to harm one to save many even when the harm is both emotionally aversive and intentional (Ciaramelli et al., 2007; Koenigs et al., 2007). We hypothesize that, in our experiment, the RTPJ was specifically recruited for reasoning about the protagonist's thoughts and knowledge, whereas the VMPFC was involved in representing the protagonist's emotionally salient, malicious desires and intentions to do harm, leading to the respective correlations with moral judgments of accidental and attempted harms. These hypotheses remain to be tested in future studies.

Second, moral judgments of accidents and attempts may depend on *quantitatively* different mental state attributions. Exculpating an agent for causing harm accidentally may require a more robust representation of the agent's belief than blaming an agent for a failed attempt. In judging an accidental harm, participants must use belief information to override a pre-potent negative response to the actual harm (Young et al., 2007). In the case of attempted harms, the outcome is neutral, so there is no salient information competing with the belief. Indeed, in the case of attempted harms, it is the mental state that is salient, insofar as the stated belief information supports further inferences of malicious desires and intentions. Thus, in development, children first use mental state information to assign blame for attempted harms, and only later learn to mitigate blame for accidents (e.g., Baird & Astington, 2004; Saxe, Carey, & Kanwisher, 2004). The strong correlation between RTPJ activation and exculpation for accidents may therefore reflect participants with especially robust belief representations. Conversely, children's relative difficulty with exculpation may be partially due to insufficiently robust mental state representations. Consistent with this hypothesis, recent research suggests the RTPJ may be late maturing (cf. Blakemore, 2008; Gogtay et al., 2004). In particular, the func-

Author's personal copy

tional selectivity of the RTPJ for beliefs increases over age, between 6 and 11 years (Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, in press). We note, though, that for both children and adults, moral exculpation likely depends not only on the capacity for mental state reasoning but also on the capacity for cognitive control. Previously, we suggested that regions for cognitive control were recruited more robustly for accidental harm than for intentional harm (Young et al., 2007), though this pattern did not replicate in the current study.

Finally, moral judgments of accidents and attempts may depend on temporally different mental state attributions. There is some evidence that consideration of negative mental states, in attempted harms, is extended over time and continues even after participants deliver their negative moral judgments (Kliemann, Young, Scholz, & Saxe, 2008; Knobe, 2005), perhaps in an effort to further support the negative moral judgments. If so, then the corresponding neural response to attempted harms may be blurred over time, and not reliably located in the tight time window surrounding the participant's button-press.

8. Other neural and cognitive processes

The predicted correlation between exculpatory moral judgment and neural response was observed in the RTPJ. This result is consistent with prior research suggesting that while other regions, including the LTPJ and MPFC support moral cognition (e.g., Greene et al., 2004) and social cognition (e.g., Mitchell, Macrae, & Banaji, 2006; Saxe & Wexler, 2005), the RTPJ may be more selective for representing beliefs both in non-moral contexts for the purpose of predicting and explaining behavior (e.g., Saxe & Powell, 2006; Perner et al., 2006) and for moral judgment (Young & Saxe, 2008). The precise role of the LTPJ and MPFC in judgment of moral scenarios, which vary the agent's intention and the action's outcome, is the topic of some of our previous and future research (Young & Saxe, 2008) and a number of studies on moral psychology (for a review, see Young & Koenigs, 2007). Undoubtedly, other brain regions (e.g., LTPJ), other cognitive processes (e.g., cognitive control, emotional empathy, emotion regulation), and other specific mental state factors (e.g., desires, goals, intentions) also contribute to moral judgment (Farrow et al., 2001; Heekeren, Wartenburger, Schmidt, Schwintowski, & Villringer, 2003; Koenigs & Tranel, 2007; Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005). These contributions, including especially the role of the VMPFC in representing emotionally salient goals, desires, and intentions for moral judgments of failed attempts to harm, should be investigated in future research.

The current results reveal a specific role for the RTPJ in processing agents' beliefs that they will do no harm. In criminal law, mistakes of fact may lead to mitigating the sentence. For future research, the law also offers a highly detailed model of the kinds and conditions of exculpation (Mikhail, 2007). First, even in the case of mistakes of fact, the mistake must be deemed "reasonable". In ongoing work, we are investigating whether the RTPJ is involved in representing not only the content, but also the reasonableness, of exculpatory beliefs. Second, the law allows for many other defenses, beyond mistakes of fact. For example, the defendant may claim to have been acting in self-defense, or to have been provoked. In general, we expect that different brain regions and not necessarily the RTPJ will be correlated with consideration of these defenses, since they do not depend directly on establishing the agent's mental states. However, these distinctions are not straightforward (e.g., the defendant must have reasonably believed that he was threatened). Finally, the law makes a distinction between two kinds of mistake: mistakes of fact (e.g., not knowing the powder was poison, which is exculpatory) and mistakes of law (e.g., not knowing it is wrong to poison someone, which is not). In future work, we will investigate whether this legal distinction corresponds to a neural division.

9. Conclusions

In sum, different levels of activation in a specific brain region for mental state reasoning, the RTPJ, track with individual differences in exculpation. Moral judgment therefore depends not just on domain-general mechanisms for abstract reasoning, cognitive control, and emotional responding, but also on distinct neural substrates for interpreting the minds of moral agents. The results may have implications for normative models of moral cognition and theory of mind, as well as for neurodevelopmental disorders such as autism that are characterized by theory of mind impairments (Blair, 1996; Leslie, Mallon, & DiCorcia, 2006). Future research in this area may also contribute to our broader understanding of forgiveness and punishment.

Acknowledgments

Many thanks to Josh Greene, Susan Carey, Fiery Cushman, and Jason Mitchell for their helpful comments, and Jon Scholz for his technical support. This project was supported by the Athinoula A. Martinos Center for Biomedical Imaging. L.Y. was supported by the NSF. R.S. was supported by MIT and the John Merck Scholars program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuropsychologia.2009.03.020.

References

Aichorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2005). Do visual perspective tasks need theory of mind? *Neuroimage*, *30*, 1059–1068.
Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development*, *103*, 37–49.
Baird, J. A., & Moses, L. J. (2001). Do preschoolers appreciate that identical actions may be motivated by different intentions? *Journal of Cognition and Development*, *2*, 413–448.
Blair, R. J. (1996). Brief report: Morality in the autistic child. *Journal of Autism and Developmental Disorders*, *26*, 571–579.
Blakemore, S. J. (2008). The social brain in adolescence. *Nature Reviews Neuroscience*, *9*, 267–277.
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.
Cacioppo, J., Petty, R., & Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307.
Chen, P., & Popovich, P. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage Publications.
Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *2*, 84–92.
Ciaramidaro, A., Adenzato, M., Enrici, I., Erk, S., Pia, L., Bara, B. G., et al. (2007). The intentional network: How the brain reads varieties of intentions. *Neuropsychologia*, *45*(13), 3105–3113.
Cohen, A. B., & Rozin, P. (2001). Religion and the morality of mentality. *Journal of Personality and Social Psychology*, *81*, 697–710.
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition*, *108*, 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.
Darley, J. M., & Zanna, M. P. (1982). Making moral judgment. *American Scientist*, *70*, 515–521.
Epstein, S., Pacini, R., DenesRaj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, *71*, 390–405.
Farrow, T. F. D., Zheng, Y., Wilkenson, I. D., Spence, S. A., Deakin, J. F. W., Tarrier, N., et al. (2001). Investigating the functional anatomy of empathy and forgiveness. *Neuroreport*, *12*, 2433–2438.

Author's personal copy

Fincham, F. D., & Jaspers, J. (1979). Attribution of responsibility to the self and other in children and adults. *Journal of Personality and Social Psychology*, *37*, 1589–1602.

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

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

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, *19*(11), 1803–1814.

Gogtay, N., Giedd, J. M., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences*, *101*, 8174–8179.

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.

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*(2), 389–400.

Grueneich, R. (1982). The development of childrens integration rules for making moral judgments. *Child Development*, *53*, 887–894.

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

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

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

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. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science*, *320*, 1092–1095.

Karniol, R. (1978). Children's use of intention cues in evaluating behavior. *Psychological Bulletin*, *85*, 76–85.

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

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

Koenigs, M., & Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: Evidence from the ultimatum game. *The Journal of Neuroscience*, *27*, 951–956.

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

Leslie, A. M., Mallon, R., & DiCorcia, J. A. (2006). Transgressors, victims, and cry babies: Is basic moral judgment spared in autism? *Social Neuroscience*, *1*, 270–283.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language*, *23*, 165–189.

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

Mikhail, J. (2007). Universal moral Grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences*, *11*, 143–153.

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

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, *6*, 799–809.

Moore, A., Clark, B., & Kane, M. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*, 549–557.

Nichols, S., & Ulatowski, J. (2007).Intuitions and individual differences: The Knobe effect revisited. *Mind and Language*, *22*, 346–365.

O'Neill, P., & Petrinovich, L. (1998). A preliminary cross-cultural study of moral intuitions. *Evolution and Human Behavior*, *19*, 349–367.

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

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology*, *64*, 467–478.

Piaget, J. (1965/1932). *The moral judgment of the child*. New York: Free Press.

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

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

Sargent, M. (2004). Less thought, more punishment: Need for cognition predicts support for punitive responses to crime. *Personality and Social Psychology Bulletin*, *30*, 1485–1493.

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and Conquer: A defense of functional localizers. *Neuroimage*, *30*(4), 1088–1096.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87–124.

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.

Saxe, R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, F. (in press). The development of brain regions for perceiving and reasoning about other people. *Child Development*.

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development*, *57*, 177–184.

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

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: Guilford Press.

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*, *104*(20), 8235–8240.

Young, L., & Koenigs, M. (2007). Investigating emotion in moral cognition: A review of evidence from functional neuroimaging and neuropsychology. *British Medical Bulletin*, *84*, 67–79.

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

Young, L., & Saxe, R. (in press). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*.

Yuill, N. (1984). Young children's coordination of motive and outcome in judgements of satisfaction and morality. *British Journal of Developmental Psychology*, *2*, 73–81.

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actors responsibility and recipients emotional reaction. *Developmental Psychology*, *24*, 358–365.

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