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

**A R T I C L E   I N F O**

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

**A B S T R A C T**

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 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.

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.

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

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;
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 makes a false belief about a white substance in her coffee. Here, the agent believes falsely that her action will be harmless, and it is her false belief that actually makes her friend sick in the end. 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, 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):

### Background
Grace and her friend are taking a tour of a chemical plant. When Grace goes over to the coffee machine to pour some coffee, Grace’s friend asks for some sugar in hers. There is white powder in a container by the coffee.

### Foreshadow
#### Negative

- The white powder is a poison left behind by a scientist.

#### Neutral

- The white powder is regular sugar left by the kitchen staff.

### Belief
#### Negative

- The container is labeled “poison”, so Grace believes that the white powder is a poison.

#### Neutral

- The container is labeled “sugar”, so Grace believes that the white powder is regular sugar.

### Outcome
#### Negative

- Grace puts the substance in her friend’s coffee. Her friend drinks the coffee and gets sick.

#### Neutral

- Grace puts the substance in her friend’s coffee. Her friend drinks the coffee and is fine.

### Judgment

Determine whether the protagonist’s belief is true. How much blame does Grace deserve for putting the substance in?

None 1 - 2 - 3 - 4 A lot

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.
(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, 13) = 0.17 p = 0.68, partial η² = 0.007); scenarios featuring negative outcomes contained the same number of words as scenarios featuring neutral outcomes (F(1, 13) = 0.17 p = 0.68, partial η² = 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 condition in each subject (Poldrack, 2006). This value was used in all analyses reported below.

### 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

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

<table>
<thead>
<tr>
<th></th>
<th>All-neutral</th>
<th>Accidental harm</th>
<th>Attempted harm</th>
<th>Intentional harm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moral judgment</td>
<td>1.5 (0.6)</td>
<td>2.1 (0.7)</td>
<td>2.9 (0.5)</td>
<td>3.7 (0.3)</td>
</tr>
<tr>
<td>Reaction time</td>
<td>2.6 (0.5)</td>
<td>2.4 (0.4)</td>
<td>2.6 (0.5)</td>
<td>2.1 (0.4)</td>
</tr>
</tbody>
</table>

Average moral judgments and reaction times (mean, standard deviation) for each of the four experimental conditions.
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).
Correlations between average percent signal (PSC) in each ROI in each condition.

<table>
<thead>
<tr>
<th></th>
<th>All-neutral</th>
<th>Accidental harm</th>
<th>Attempted harm</th>
<th>Intentional harm</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTPJ</td>
<td>r(15) = −0.38, p = 0.17</td>
<td>r(15) = −0.66, p = 0.807</td>
<td>r(15) = 0.19, p = 0.50</td>
<td>r(15) = −0.30, p = 0.28</td>
</tr>
<tr>
<td>PC</td>
<td>r(15) = −0.06, p = 0.82</td>
<td>r(15) = −0.14, p = 0.82</td>
<td>r(15) = −0.24, p = 0.38</td>
<td>r(15) = −0.46, p = 0.08</td>
</tr>
<tr>
<td>LTPJ</td>
<td>r(15) = −0.02, p = 0.94</td>
<td>r(15) = −0.19, p = 0.49</td>
<td>r(15) = 0.35, p = 0.20</td>
<td>r(15) = −0.38, p = 0.52</td>
</tr>
<tr>
<td>DMPC</td>
<td>r(13) = −0.02, p = 0.96</td>
<td>r(13) = 0.18, p = 0.56</td>
<td>r(13) = −0.28, p = 0.36</td>
<td>r(13) = −0.41, p = 0.65</td>
</tr>
<tr>
<td>MMPFC</td>
<td>r(10) = −0.55, p = 0.10</td>
<td>r(10) = 0.07, p = 0.85</td>
<td>r(10) = 0.52, p = 0.12</td>
<td>r(10) = 0.17, p = 0.65</td>
</tr>
<tr>
<td>VMPC</td>
<td>r(10) = 0.26, p = 0.047</td>
<td>r(10) = −0.49, p = 0.15</td>
<td>r(10) = −0.64, p = 0.048</td>
<td>r(10) = −0.51, p = 0.13</td>
</tr>
</tbody>
</table>

Significant (p < 0.05) values bolded.
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, Denes-Raj, & 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” put 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, Mucchioli, 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.

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 & Astinong, 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 insufficiency robust mental state representations. Consistent with this hypothesis, recent research suggests the RTPJ may be late maturing (cf. Blakemore, 2008; Cogtay et al., 2004). In particular, the func-
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, Schiwintoski, & 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, & DiCicco, 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


