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.

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

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):

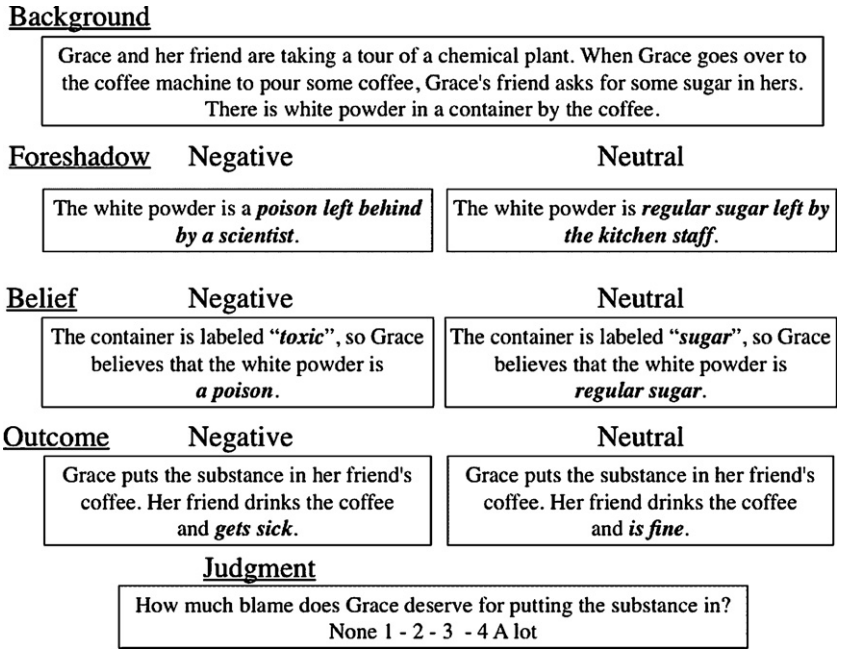

Fig. 1. Experimental stimuli and design. "Foreshadow" information foreshadows whether the action will result in a negative or neutral outcome. "Belief" information states whether the protagonist holds a belief that she is in a negative situation and that action will result in a negative outcome ("negative" 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.