Check for updates

Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation

Jana Schaich Borg, Catherine Hynes, John Van Horn, Scott Grafton, and Walter Sinnott-Armstrong

Downloaded from http://direct.mit.edu/jocn/article-pdf/18/5/803/1958540/jocn.2006.18.5.803.pdf by guest on 18 May 2021

Abstract

■ The traditional philosophical doctrines of Consequentialism, Doing and Allowing, and Double Effect prescribe that moral judgments and decisions should be based on consequences, action (as opposed to inaction), and intention. This study uses functional magnetic resonance imaging to investigate how these three factors affect brain processes associated with moral judgments. We find the following: (1) Moral scenarios involving only a choice between consequences with different amounts of harm elicit activity in similar areas of the brain as analogous nonmoral scenarios; (2) Compared to analogous nonmoral scenarios, moral scenarios in which action and inaction result in the same amount of harm elicit more activity in areas associated with cognition (such as the dorsolateral prefrontal cortex) and less activity in areas associated with emotion (such as the orbitofrontal cortex and temporal pole); (3) Compared to analogous nonmoral scenarios, conflicts between goals of minimizing harm and of refraining from harmful action elicit more activity in areas associated with emotion (orbitofrontal cortex and temporal pole) and less activity in areas associated with cognition (including the angular gyrus and superior frontal gyrus); (4) Compared to moral scenarios involving only unintentional harm, moral scenarios involving intentional harm elicit more activity in areas associated with emotion (orbitofrontal cortex and temporal pole) and less activity in areas associated with cognition (including the angular gyrus and superior frontal gyrus). These findings suggest that different kinds of moral judgment are preferentially supported by distinguishable brain systems. ■

INTRODUCTION

Ever since Socrates debated sophists and stoics wrangled with skeptics, philosophers have argued about whether moral judgments are based on reason or on emotion. Although definitions of morality may vary across cultures and philosophies in other ways, all definitions of ''moral judgments'' include judgments of the rightness or wrongness of acts that knowingly cause harm to people other than the agent. These central moral judgments are distinct from economic or prudential judgments based on the agent's own interest, both because the moral judgments depend on interests of other people and because they focus on harms as opposed to, say, pleasure. The present study addresses core moral judgments of this kind (Nichols, 2004).

To investigate whether such core moral judgments are based on emotion or on reason, these terms must first be defined. For our purposes, ''emotions'' are immediate valenced reactions that may or may not be conscious. We will focus on emotions in the form of negative affect. In contrast, ''reason'' is neither valenced nor immediate insofar as reasoning need not incline us toward any specific feeling and combines prior information with new beliefs or conclusions and usually comes in the form of cognitive manipulations (such as evaluating alternatives) that require working memory. Emotion might still affect, or even be necessary for, reasoning (Damasio, 1994), but emotion and reasoning remain distinct components in an overall process of decision making.

In modern times, Hume (1888) and many utilitarian philosophers based morality on emotion or sentiment via what the former called ''sympathy'' and what contemporary psychologists call ''empathy.'' In their view, core moral judgments arise from an immediate aversive reaction to perceived or imagined harms to victims of actions that are judged as immoral only after and because of this emotional reaction. In contrast, Kant (1959) insisted that his basic nonutilitarian moral principle (the categorical imperative) could be justified by pure reason alone, and particular judgments could then be reached by reasoning from his basic principle, all without any help from emotion. Although somewhat transformed, this fundamental debate still rages among philosophers today.

Such traditional issues are difficult to settle in an armchair, yet some progress has been made with the help of recent brain imaging techniques. Studies using

Dartmouth College

© 2006 Massachusetts Institute of Technology

functional magnetic resonance imaging (fMRI) surprisingly suggest that neither Kant nor Hume had the whole truth, and that some moral judgments involve more emotion whereas others involve more reasoning. For example, neural systems associated with emotions are activated more by personal moral dilemmas than by impersonal moral dilemmas (Greene, Nystrom, Engell, & Darley, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Personal moral dilemmas were labeled ''personal'' because the agent gets ''up close and personal'' with the victim in most such cases. Greene, Nystrom, et al. (2004) and Greene, Sommerville, et al. (2001) formally define personal moral dilemmas, without reference to physical proximity, as cases where an otherwise desirable action is (1) likely to cause serious bodily harm (2) to a particular person or group of persons (3) not by deflecting an existing threat onto a different party. A paradigm personal moral dilemma is the footbridge case, where the only way to save five people from a runaway trolley is to push a fat man off of a footbridge in front of the trolley so as to stop the trolley before it hits the five people. A paradigm impersonal moral dilemma is the sidetrack case, where the only way to save five people is to redirect a runaway trolley onto a sidetrack where it will kill one person. Most people judge that it is morally wrong to push the fat man off the footbridge; fewer people judge that it is morally wrong to redirect the trolley onto the sidetrack. The question is why and how people make these contrasting moral judgments.

These trolley cases differ as to whether a threat is created (as in pushing the fat man) or merely deflected (as in redirecting the trolley onto the sidetrack), but there are other differences as well. For example, the agent gets closer to the victim in the footbridge case and that the agent in the footbridge case could jump in front of the trolley instead of pushing the fat man, whereas this option is not available in the sidetrack case. Such complications presumably explain why Greene, Sommerville, et al. (2001) admit that their distinction between personal and impersonal scenarios is only ''a useful 'first cut,' an important but preliminary step toward identifying the psychologically essential features of circumstances that engage (or fail to engage) our emotions and that ultimately shape our moral judgments . . .'' (p. 2107).

The exploratory fMRI study reported here makes a ''second cut'' by evaluating three factors picked out by traditional moral principles that might underlie the distinction between personal and impersonal moral problems. One classic moral theory is Consequentialism, which claims roughly that we morally ought to do whatever has the best consequences overall (Sinnott-Armstrong, 2003). Opposed to Consequentialism are two deontological principles. The Doctrine of Doing and Allowing (DDA) says that it takes more to justify doing harm than to justify allowing harm; thus, for example, it is sometimes morally wrong to commit an act of killing in circumstances where it would not be morally wrong to let someone die by refraining from an act of saving (Howard-Snyder, 2002). Redirecting the trolley onto the sidetrack violates the DDA insofar as it involves positive action that causes death and, thus, counts as killing, but the DDA is not violated by merely letting someone drown to be able to save five other drowning people. The Doctrine of Double Effect (DDE), in contrast, holds that it takes more to justify harms that were intended either as ends or as means than to justify harms that were known but unintended side effects; thus, for example, it is sometimes morally wrong to intend death as a means when it would not be morally wrong to cause that death as an unintended side effect (McIntyre, 2004). Pushing the fat man in front of the trolley in the footbridge case violates the DDE because the agent intends to use the fat man as a means to stop the trolley and save the five people on the main track. In contrast, deflecting the trolley in the sidetrack case does not violate the DDE because the victim's death in that case is only an unintended side effect that is not necessary for the agent's plan to succeed in saving the five people on the main track. Both the DDA and the DDE conflict with Consequentialism because these deontological principles claim that factors other than consequences matter, so it is sometimes morally wrong to do what has the best consequences overall.

An empirical study cannot, of course, determine which moral theory is correct or which acts are morally wrong. That is not the purpose of our study. Our goal is only to use traditional theories, which pick out factors that affect many people's moral intuitions, as tools to explore neural systems involved in moral judgment. The factors used in this study play a significant role not only in moral philosophy but also in law (because people are usually not guilty of first-degree murder when they merely let people die and do not intend death) and religion (such as when the Catholic Church cites a prohibition on intended harm to justify its official positions on abortion and euthanasia). Psychological studies have documented omission bias (Kahneman & Tversky, 1982) and intention bias (Hauser, 2006) in moral judgment, substantiating the impact of action and intention on law and religion. Other factors, such as proximity to the victim and creation vs. deflection of a threat, may also affect moral judgments but were not investigated explicitly in this study.

Based on previous studies of neural correlates of moral judgments (surveyed in Moll, de Oliveira-Souza, & Eslinger, 2003; Greene & Haidt, 2002), we hypothesized that: (1) the medial frontal gyrus (Brodmann's area [BA] 10), (2) the frontopolar gyrus (BA 10), and (3) the posterior superior temporal sulcus (STS)/inferior parietal lobe (BA 39) would be more active when considering moral scenarios than when considering nonmoral scenarios, irrespective of consequences, action, and inten-

Downloaded from http://direct.mit.edu/jocn/article-pdf/18/5/803/1958840/jocn.2007.18.5.803.pdf by guest on 18 May 2021

tion. Hypotheses regarding the differential effects of consequences, action, and intention were then framed with respect to anatomic circuits linked to emotion and cognition. The paralimbic system, including the amygdala, cingulate cortex, hippocampal formation, temporal pole, and ventromedial prefrontal (including orbitofrontal) cortex (Mesulam, 2000), has been credited as the ''emotional'' system of the brain, whereas the ''central executive'' system (Baddeley, 1986), including regions of the parietal lobe and lateral regions of the prefrontal cortex, has been credited as the ''cognitive'' system of the brain (Miller & Cohen, 2001). We hypothesized that negative affect or emotion would be associated with violations of established moral doctrines, and thus (4) the paralimbic system would be activated more in thinking about actions that cause harm (and thus violate the DDA) than in thinking about similarly harmful nonactions or action omissions. We further hypothesized (5) that the paralimbic system would be more activated when thinking about intending harm as a means (which violates the DDE) than in thinking about causing harm as an unintended side effect. Conversely, cognitive regions of the brain involved in reasoning would be activated relatively more in considering moral scenarios that did not include violations of either the DDA or the DDE.

Another factor that is often said to affect emotion and moral judgment is the language used to describe a scenario. Past studies of moral judgment seem to describe moral or personal moral scenarios with more colorful language than their nonmoral or ''impersonal'' moral counterparts. (Greene Nystrom, et al., 2004; Greene, Sommerville, et al., 2001) If so, this confound might explain greater activations observed in emotional systems. To test the effects of language, we presented each moral scenario in both dramatic (colorful) and muted (noncolorful) languages. We hypothesized (6) that moral scenarios presented in a colorful language would activate regions of the paralimbic system more than otherwise similar moral scenarios presented in a plain language.

The experimental design used to test these hypotheses required some unique features. Previous studies on moral judgment compared moral stimuli to non-moral unpleasant stimuli (Moll, de Oliveira-Souza, Bramati, & Grafman, 2002) or semantic improprieties (Heekeren, Wartenburger, Schmidt, Schwintowski, & Villringer, 2003), neither of which consistently involve nonmoral social processes. As a result, activations that appear in their respective moral vs. control contrasts may represent general social processing rather than uniquely moral processing. Furthermore, Greene, Nystrom, et al. (2004), Heekeren et al. (2003), and Greene, Sommerville, et al. (2001) ambiguously asked their subjects to judge whether actions in their moral conditions were ''appropriate'' or ''inappropriate.'' It is unclear how subjects construed this request (according to their own moral values, what society deems acceptable, or what is legal), making it difficult to determine whether the aforementioned study results really reflect the processes that underlie moral judgment in particular. Moreover, the cognitive processing required by previous control conditions was only weakly matched to that required by their moral conditions, again making it difficult to determine which cognitive processes accompany moral judgment in comparison to other kinds of social judgment.

To avoid possible confounds of past studies, we restricted our moral scenarios to issues of killing and letting die rather than other moral topics, such as rape, theft, and lying. Our nonmoral scenarios described destruction of objects of personal value rather than harm to other people. Hence, although our nonmoral scenarios involved other people (such as firefighters and clerks) and drew upon other kinds of social processing, they did not require any core moral judgments or specifically moral processing. All variables of the factorial design were matched so that nonmoral scenarios had the same combinations of consequence, action, intention, and language conditions as moral scenarios. Moral scenarios were then compared directly to nonmoral scenarios, rather than to a baseline or a separate less demanding cognitive condition. Instead of asking the subjects whether it would be appropriate to perform an action, we asked ''Is it wrong to (action appropriate to the scenario)?'' and ''Would you (action appropriate to the scenario)?'' By asking both questions, we hoped to reduce the risk that different subjects would approach the scenarios with different questions in mind.

METHODS

Experimental Design

The factors described in the introduction were operationalized into four variables (morality, type, means, and language; Table 1) and entered into a factorial design (Table 2).

The morality variable had two levels: ''Moral'' scenarios described harm to people, and ''nonmoral'' scenarios described harm to objects of personal value. Thus, only moral scenarios asked for core moral judgments as defined in the Introduction. The type variable had three levels: ''numerical consequences,'' ''action,'' and ''both.'' ''Numerical consequences'' scenarios described an action that would harm a smaller number of people/objects and another action that would harm a larger number of people/objects. Because it would be nonsensical to offer separate options describing two different inactions, options in numerical consequences scenarios were presented as positive actions. More harmful options represented violations of Consequentialism (although consequentialists take into consideration many

Downloaded from http://direct.mit.edu/jocn/article-pdf/20/5/804/1958840/jocn.2008.20057.pdf by guest on 18 May 2021

Table 1. Experimental Variables

Variable	*Level*	*Description*
Morality	Moral	Acting on people (e.g., trolley scenario)
	Nonmoral	Acting on objects and possessions
Type	Numerical consequences	Harming *x* people/objects vs. harming *y* people/objects (Consequentialism)
	Action	Harming *x* people/objects vs. letting *x* people/objects be harmed (DDA)
	Both (numerical consequences + action)	Harming *x* people/objects vs. letting *y* people/objects be harmed (Consequentialism × DDA)
Means	Means	Intentionally using some people/objects as a means to save others (DDE)
	Nonmeans	Causing unintentional but foreseen harm to people/things to save others
Language	Colorful	Described with more detailed imagery and dramatic words
	Plain	Described with plain imagery and simple words

DDA = Doctrine of Doing and Allowing; DDE = Doctrine of Double Effect.

effects other than the number harmed). ''Action'' scenarios then described an action that would harm the same number (but a different group) of people/objects as would be harmed if the act were omitted. Action scenarios, therefore, proposed violations of the DDA. ''Both'' scenarios described an action that would harm fewer people than would be harmed if the act were omitted, thus portraying conflicts between the DDA and Consequentialism. If numerical consequences and action had been separated into independent two-level variables, one of their interactions would have been a cell describing two options of action that saved/killed the same number of people. Given that all parties in the scenarios were anonymous and that all other variables were held constant, subjects would have had to choose arbitrarily. Because we would have had no way to control for the influences on such arbitrary choices and because the motivations behind such choices would likely involve nonmoral processing, we combined the numerical consequences and action variables into the three-level variable, ''type.''

The means variable had two levels: ''means'' scenarios, which described intended harm, and ''nonmeans'' scenarios, which described foreseen but unintended harm. Means scenarios proposed violations of the DDE.

The language variable also had two levels: ''Colorful'' scenarios were described in dramatic language, and ''plain'' scenarios were described in muted language.

Our four variables together constituted a 2 (Morality) × 3 (Type) × 2 (Means) × 2 (Language) design (Table 2). Due to timing constraints, we had two scenarios in each of the moral factor cells (24 moral scenarios) and one scenario in each of the nonmoral factor cells (12 nonmoral scenarios).

Each scenario block consisted of a series of three screens (Figure 1). The first screen described the scenario. The second and third screens posed the questions ''Is it wrong to (action appropriate to the scenario)?'' and ''Would you (action appropriate to the scenario)?'', which were presented in randomized order. Subjects read and responded to the scenarios at their own pace, pressing the right button to answer ''yes'' and the left button to answer ''no.'' Each response advanced the stimulus program to the next screen. Subjects' responses and response times to both questions were recorded. Four runs of nine fully randomized scenario blocks were presented with 30 sec of rest at the beginning and at the end of each run. A presentation software (http://nbs.neuro-bs.com) was used for presenting all stimuli and for recording responses and response times.

Subjects were informed of the provocative nature of the scenarios before entering the scanner. They were also told that they would have to answer the questions ''Is it wrong to (action appropriate to the scenario)?'' and ''Would you (action appropriate to the scenario)?'' after each scenario, so that they would understand the

Table 2. Experimental Design

Variable	*Factorial Design*																							
Morality	Moral												Nonmoral											
Type	Numerical consequences				Action				Both				Numerical consequences				Action				Both			
Means	Means		Nonmeans		Means		Nonmeans		Means		Nonmeans		Means		Nonmeans		Means		Nonmeans		Means		Nonmeans	
Language	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C

Downloaded from http://direct.mit.edu/jocn/article-pdf/18/5/793/1958840/jocn.2006.18.5.793.pdf by guest on 18 May 2021

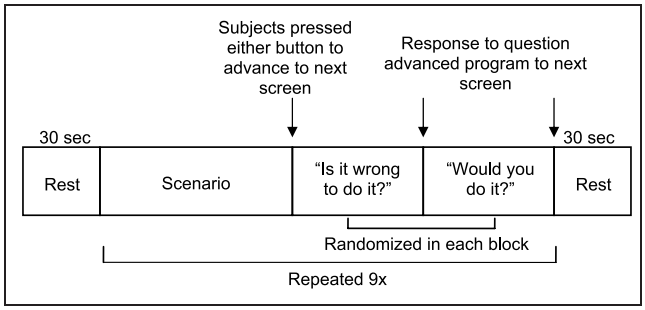

Figure 1. fMRI functional run.

difference between the two types of questions before the scans were collected.

Stimuli

We developed 36 scenarios to fit into the 2 × 3 × 2 × 2 factorial design representing the variables morality, type, means, and language, respectively. The moral scenarios described variations of moral dilemmas frequently discussed in moral philosophical literature (much of which are collected in Fischer & Ravizza, 1992). A full list of the experimental scenarios can be found at http://dbic.dartmouth.edu/~borg. Some examples are as follows:

Moral/action/nonmeans/plain language: *You find yourself locked in a room in front of a one-way mirror. On the other side of the mirror, there is a machine holding six people, three on each side. A sign on the floor says that if you do nothing, in 10 sec, the three people on the left will be shot and killed and the three people on the right will live. If you push the button in front of you, in 10 sec, the three people on the right will be shot and killed, but the three people on the left will live. There is no way to prevent the deaths of both groups of people. Is it wrong to push the button so that the three people on the right will be shot? Would you push the button so that the three people on the right will be shot?*

Nonmoral/both/means/colorful language: *The extremely hot and parching weather has started a roaring brush fire next to your property, which has been painstakingly sown with wildflowers. If you do nothing, the emergency team will come in their yellow fireproof suits to put out the fire with water—5 acres of your land will be charred to the ground before the fire is extinguished. If you purposely burn down two acres of your wildflower land at the edge of your property in a safe controlled manner as a firebreak, the fire will put itself out when it reaches your property boundaries without destroying any more of your wildflower land. There is no other way to prevent damage to both pieces of your property. Is it wrong to burn the 2-acre firebreak in your wildflower land? Would you burn the 2-acre firebreak in your wildflower land?*

The scenarios were standardized for length and semantic complexity. All scenarios were scored for content with respect to the experimental variables by three independent evaluators who had been given prior training in the meaning of the experimental variables used. With one exception by one evaluator, all scenarios were placed in the intended cell of the factorial design by all three evaluators. In a parallel behavioral study of 51 subjects (results to be discussed in another article), all variables (morality, type, means, and language) proved to have significant effects ($p < .05$ to $p < .0001$) on subject responses and/or response times.

Subjects

Ethics approval for the study was obtained through the Dartmouth College Committee for Protection of Human Subjects (Hanover, NH, USA). Twenty-eight participants (16 men, 12 women) were recruited from the local Dartmouth College community. One participant's data were discarded due to technical difficulties and/or data corruption, and three participants' data were discarded because these subjects took too much time to complete the experiment. The remaining 24 participants (13 men, 11 women) were included in the analysis. The mean age of the subjects was 22.9 years, and ages ranged from 18 to 32 years. All subjects were right-handed on self-report.

Imaging

Blood-oxygen-level-dependent (BOLD) functional imaging was obtained with an ultrafast echo planar gradient imaging sequence: T_R = 2.5 sec, T_E = 35 sec, flip angle = 90°, twenty-five 4.5-mm interleaved slices separated by a 1-mm gap per T_R, in-plane resolution = 3.125 × 3.125 mm, BOLD images tilted 15° clockwise from the AC–PC plane. A typical run was about 9 min long. A set of corresponding anatomical images was obtained with a T1-weighted axial fast-spin echo sequence, with 25 contiguous slices coplanar to BOLD images, slice thickness = 4.5 mm, gap = 1.0 mm, T_E = min full, T_R = 650 msec, echo train = 2, field of view = 24 cm. In addition, high-resolution (0.94 × 0.94 × 1.2 mm) whole-brain 3-D spoiled gradient recall acquisition (SPGR) structural images were collected. All images were obtained with a 1.5-T General Electric Horizon whole-body MRI scanner, using a standard birdcage headcoil (GE, Milwaukee, WI).

Downloaded from http://direct.mit.edu/jocn/article-pdf/18/5/807/1958540/jocn.2008.20057.pdf by guest on 18 May 2021

Preprocessing was performed using SPM99 software (http://www.fil.ion.ucl.ac.uk/spm).

Functional and structural images were coregistered and transformed into a standardized reference template from the Montreal Neurological Institute (Montreal, Quebec, Canada). The final voxel size was $3 \times 3 \times 3$ mm^3, and a 10-mm smoothing kernel was applied to the data.

Group Statistical Analysis

All group analyses were performed using custom-designed programs in MATLAB (The Mathworks, Natick, MA; version 6.0.3). To maximize the detection of commonly activated regions, we entered raw signal intensity values into the group analyses without any individual subject or omnibus statistical thresholding. Instead, a custom-made mask defined by the mean of all functional scans across subjects was applied to restrict statistical analysis to only those voxels where data were consistently present in the group. Before being entered, the data of each voxel were globally normalized across time to remove variance associated with scanner differences from subject to subject. Normalized image volumes associated with reading and responding (from the moment they started reading the scenario to the moment they answered the second ''Is it wrong to (action appropriate to the scenario)?'' or ''Would you (action appropriate to the scenario)?'' question) to scenarios in each analysis of variance (ANOVA) cell were averaged to reduce within-trial variability. Event times were offset by 5 sec to account for hemodynamic response function. The consolidated globally normalized signal temporally adjusted for hemodynamic delay was subsequently entered into a $2 \times 3 \times 2 \times 2$ mixed-effects ANOVA, with experimental variables entered as fixed effects and with subjects entered as random effect. Significant regions of interest were defined as clusters of 10 voxels or more with F values of $p < .05$. Because the regions identified were consistent with a priori predictions of sites engaged in moral reasoning, as defined by prior studies (Greene & Haidt, 2002), correction for multiple comparisons over the entire brain volume was not performed. Significant sites were superimposed on a high-resolution scan from one of the subjects for anatomic visualization. Each site was localized with respect to the anterior commissure and converted to Talairach coordinates using the Talairach demon (Talairach & Tournoux, 1988).

RESULTS AND DISCUSSION

Response choice (Table 3), reaction time data (Table 3), and imaging results (representing brain activity during the reading and response phases combined; Table 4) confirm that Consequentialism, the DDA, the DDE, and language affected the responses of our fMRI subject population. Brain and relevant behavioral data will be discussed for the main effect of morality and for each variable, respectively.

Main Effect of Morality

Manipulation of morality was successful in modulating subjects' responses to scenarios that were equal in respect to all other experimental variables. Subjects' choices indicated that they were more willing to act upon objects of personal value than to act upon people, irrespective of resulting consequences or motivating intentions of the act, and irrespective of the language used to describe the act (see Table 3 for specific behavioral data). Nevertheless, reaction times to the question ''Is it wrong to do it?'' did not differ significantly ($p = .35$) between the moral and the nonmoral conditions, suggesting that the two conditions were well-matched for complexity and difficulty. Subjects answered ''Would you do it?'' more quickly in moral scenarios than in nonmoral scenarios ($p = .002$) however, suggesting that decisions about what to do are processed differently than judgments about what it is wrong to do. This indicates that future studies of morality should distinguish between the two questions.

We predicted that the medial frontal gyrus (BA 10), frontal pole (BA 10), and STS/inferior parietal lobe (BA 39) should be more activated in moral scenarios than in nonmoral scenarios, even after adjusting for the variance of Type, Means, and Language. The medial frontal gyrus (rostral gyrus, 105 voxels, BA 10) was activated more in moral scenarios than in nonmoral scenarios (Table 4, Figure 2), supporting our first hypothesis. The local maximum observed in this study is more ventral than the local maximum reported by Heekeren et al. (2003), Greene, Sommerville, et al. (2001), or Moll, Eslinger, and Oliveira-Souza, (2001), but the region is bilateral and extends dorsally to the areas reported by previous studies. The medial frontal gyrus is also reliably activated in self-referential processing (Johnson et al., 2002; Kelley et al., 2002). One previous paradigm investigating morality used a control condition that necessitated reference to the self (Greene, Nystrom, et al., 2004; Greene, Sommerville, et al., 2001). Our results suggest that moral judgment may utilize more self-referential processing than even important judgment about one's possessions or livelihood, consistent with the idea that we consider our morals to be crucially defining parts of who we are and who we want to be.

Our second hypothesis was also substantiated. We identified a frontopolar region (BA 10) in the left hemisphere that was more active in moral scenarios than in nonmoral scenarios (Table 4, Figure 2), replicating the results of Moll, Eslinger, et al. (2001), but not those of Greene, Nystrom, et al. (2004), Heekeren et al. (2003), or Greene, Sommerville, et al. (2001). The

Downloaded from http://direct.mit.edu/jocn/article-pdf/18/5/803/1958840/jocn.2006.18.5.803.pdf by guest on 18 May 2021

Table 3. Behavioral Data

Variable	*Factorial Design*											
Moral												
Type	Numerical consequences				Action				Both			
Means	Means		Nonmeans		Means		Nonmeans		Means		Nonmeans	
Language	P	C	P	C	P	C	P	C	P	C	P	C
% Yes, WRT	73	65	44	52	69	75	65	73	63	88	67	79
% Yes, WOY	38	33	63	63	8	8	13	8	40	13	46	25
RT (sec), WRT	2.54 ± 3.03	2.17 ± 2.33	2.41 ± 2.02	2.60 ± 2.38	1.81 ± 1.47	2.24 ± 2.44	2.35 ± 3.35	2.02 ± 1.81	2.02 ± 1.58	1.73 ± 1.55	2.49 ± 2.40	1.83 ± 1.67
RT (sec), WOY	2.21 ± 1.88	1.74 ± 1.50	2.27 ± 2.92	2.55 ± 2.38	1.67 ± 1.63	1.61 ± 1.60	1.71 ± 1.48	1.55 ± .97	2.14 ± 1.20	2.53 ± 3.48	2.44 ± 2.67	2.22 ± 2.34
Nonmoral												
Type	Numerical consequences				Action				Both			
Means	Means		Nonmeans		Means		Nonmeans		Means		Nonmeans	
Language	P	C	P	C	P	C	P	C	P	C	P	C
% Yes, WRT	4	8	4	8	13	54	17	8	13	17	21	13
% Yes, WOY	92	92	88	92	54	54	8	54	92	88	83	88
RT (sec), WRT	1.48 ± 1.08	1.40 ± 2.28	2.76 ± 3.02	2.61 ± 2.00	2.23 ± 2.76	4.24 ± 3.30	2.99 ± 2.81	2.22 ± 2.27	1.35 ± .79	1.37 ± 2.39	2.62 ± 2.34	2.61 ± 1.80
RT (sec), WOY	1.27 ± .68	1.37 ± 2.61	2.03 ± 3.61	3.75 ± 3.72	4.09 ± 3.28	3.34 ± 1.99	4.93 ± 3.91	3.12 ± 4.17	1.86 ± 1.66	1.44 ± 4.05	2.42 ± 1.91	2.28 ± 2.10

Percentages of yes responses and mean reaction time to the questions "Is it wrong to do it?" (WRT) and "Would you do it?" (WOY) for each cell of the experimental design are presented.

Downloaded from http://direct.mit.edu/jocn/article-pdf/16/5/807/1935589/089892904750361903.pdf by guest on 18 May 2021

Table 4. fMRI Data

Region	Brodmann's Area	MNI			k	F
		x	y	z		
Moral > nonmoral						
Bilateral inferior/superior rostral	10	−12	51	−9	105	11.65
L frontal pole	10	−15	72	−6	24	8.23
L lingual gyrus	18	−6	−72	−3	10	6.89
R temporal–occipital transition zone, pSTS	19/39	51	−75	9	82	14.22
L temporal–occipital transition zone, pSTS	19/39	−54	−72	6	16	11.82
L supramarginal gyrus	40	−63	−51	21	24	4.39
L middle frontal, caudal DLPFC	9/46	−42	48	30	15	5.78
Nonmoral > moral						
R middle frontal gyrus, caudal DLPFC	46	48	45	27	11	5.82
L medial superior frontal gyrus	8	−3	21	69	17	5.69
Morality × Type						
L middle temporal gyrus	38	−57	12	−33	18	5.75
L lateral orbital frontal gyrus	11	−27	45	−12	149	4.98
R middle frontal gyrus, rostral DLPFC	8	57	9	36	491	8.6
L middle occipital gyrus	18	−27	−84	−6	24	4.55
L posterior inferior temporal gyrus	37	−48	−63	−6	18	4.24
L supramarginal gyrus	40	−60	−33	33	96	6.83
R supramarginal gyrus	40	63	−24	27	42	6.6
L angular gyrus	40	−39	−78	45	10	3.63
R central sulcus	4	42	−15	63	24	4.74
Bilateral medial superior frontal gyri, SMA	6	0	6	75	84	8.53
Morality × Means						
R anterior superior temporal sulcus, STS	21/22	63	−9	−18	18	6.94
L inferior rostral gyrus	11	−6	63	−15	32	8.47
R frontal pole	10	30	66	−9	41	10.01
R calcerine sulcus	17	6	−81	18	27	5.27
R superior occipital gyrus	19	21	−99	21	48	11.08
R supramarginal gyrus	40	63	−45	36	19	6.83
R angular gyrus	40	51	−63	54	21	5.11
Medial SMA	6	3	−15	75	21	6.65
Morality × Language						
R entorhinal/fusiform gyrus	36/20	27	−24	−30	52	9.81
R posterior orbital frontal gyrus	11	24	30	−12	274	12.77
R lateral temporal pole	38	48	15	−12	10	5.67
Anterior cingulate cortex	32	−3	42	−3	28	7.76
R superior rostral gyrus	10	6	61	3	44	9.22
R transverse temporal sulcus	42	60	−12	3	61	7.41
R superior frontal, DLPFC	8	36	39	57	12	13.9

The Brodmann's Areas, MNI coordinates, number of voxels (*k*), and *F* value of each region of interest according to the interaction in which they appeared ($p < .05$) are presented. Reported coordinates and *F* values are for the cluster maxima.

L = left; R = right.

Downloaded from http://direct.mit.edu/jocn/article-pdf/18/5/810/1958840/jocn.2006.18.5.810.pdf by guest on 18 May 2021

frontal pole has also been implicated in moral processing through lesion studies. In fact, among all prefrontal patients, those with frontopolar lesions are often the most problematic and the most difficult to treat for sociopathic symptoms (Eslinger, Flaherty-Craig, & Benton, 2004). Some argue that sociopathic problems arise because one of the functions of the frontal pole is to maintain long-term goals while more immediate goals are being processed (Braver & Bongiolatti, 2002; Koechlin, Basso, Pietrini, Panzer, & Grafman, 1999). Mutual cooperation and altruism often require that individual moral and social decisions be made according to overarching abstract objectives, rather than immediate goals. Thus, if patients are unable to maintain and act according to these long-term goals, it is likely that they will demonstrate impaired sociomoral behavior. Another implication of this possible function of the frontal pole is that moral situations requiring the maintenance and/or comparison of multiple social goals may enlist the frontal pole more than other less discordant situations referencing fewer social goals. We presented moral dilemmas that called upon competing moral intuitions and that sometimes had no clear resolution. In contrast, Heekeren et al. used moral stimuli that are explicitly designed to be simple and easy (presumably requiring less maintenance and manipulation of competing intuitions and goals), and many of the stimuli of Greene, Nystrom, et al. and Greene, Sommerville, et al. were very easy to respond to as well (as delineated in the ''easy'' vs. ''difficult'' contrast in 2004). If the presented hypothesis of frontal pole function is correct, then this may explain why these fMRI studies did not find the frontal pole to be more active in moral decision making than in control tasks.

Supporting our third hypothesis, we found bilateral activation of the posterior STS/inferior parietal lobe (BA 19/39) in the main effect of morality (Table 4, Figure 2). Heekeren et al. (2003) and Moll, de Oliveira-Souza, Bramati, et al. (2002) found areas of BA 39 in the right and left hemispheres, respectively, to be active in their moral conditions; our activation was about 1.5 cm posterior of their reported regions. Greene, Nystrom, et al. (2004) and Greene, Sommerville, et al. (2001) found the posterior STS (BA 39) to be more active in personal moral scenarios than in impersonal moral scenarios. In contrast, anterior regions of the STS (BA 21/22) were identified only in the Morality × Means interaction, and not in the main effect of morality (inconsistent with the results of Heekeren et al., 2003; Moll, de Oliveira-Souza, Bramati, et al., 2002; Moll, de Oliviera-Souza, Eslinger, Bramati, et al., 2002).

The posterior STS (BA 19/39), a visual and auditory integration area implicated in analogy tasks (Assaf et al., in press) and other aspects of language requiring auditory short-term memory (Dronkers, Wilkins, Van Valin, Redfern, & Jaeger, 2004), has been shown to be functionally distinct from the anterior STS, which seems to be active in more automatic linguistic tasks, such as syntax detection, and is specialized to just one sensory modality (Beauchamp, Lee, Argall, & Martin, 2004). There is also evidence that the anterior STS, but not the posterior STS, is active during belief-laden reasoning (Goel & Dolan, 2003). It is likely that the posterior STS plays a correspondingly distinct role in moral processing. Accordingly, Greene, Nystrom, et al. (2004) found the posterior STS to be preferentially active in difficult personal scenarios, whereas the anterior STS was preferentially active in easy personal scenarios. One possible explanation for the observed activation in many regions of the STS along the anterior–posterior continuum is that the posterior STS may play a preferential role in thought-provoking first-time moral judgment that requires executive resources, whereas the anterior STS may be more involved in previously resolved routine moral judgment that requires more semantically based representational knowledge. More investigation is needed to determine the roles of STS regions in the current paradigm, but we underline the need for future studies to distinguish between the functions of the anterior and the posterior STS in moral judgment.

Not addressed in our experimental hypotheses, a region of the left rostral dorsolateral prefrontal cortex (DLPFC; BA 9/46) associated with high-level executive function and ''active retrieval'' in working memory (Kostopoulos & Petrides, 2003; Miller & Cohen, 2001), which is also activated during the moral conditions used by Heekeren et al. (2003) and Moll, Eslinger, et al. (2001), was found to be more active in moral judgment than in nonmoral judgment once all variances associated with other experimental variables have been accounted for. However, the *right* DLPFC was more active in nonmoral processing than in moral processing, as it was in the impersonal moral and nonmoral scenarios of Greene, Nystrom, et al. (2004) and Greene, Sommerville, et al. (2001). The left and right DLPFC regions identified in moral and nonmoral decision making, respectively, had a similar magnitude of response, as assessed by local peak *F* values. We speculate, therefore, that moral and nonmoral scenarios required equal amounts of working memory and executive cognitive function, but that they required *different types* of working memory and executive cognitive function determined by the methods used to answer and/or process the experimental questions. It has been suggested that the left DLPFC is primarily responsible for working memory and cognitive processes involved in verbal reasoning, whereas the right DLPFC is more responsible for working memory and cognitive processes involved in spatial and other forms of nonverbal reasoning (Suchan et al., 2002; Caplan & Dapretto, 2001; Chee, O'Craven, Bergida, Rosen, & Savoy, 1999). The left, but not the right, DLPFC is consistently activated in semantic association tasks requiring subjects to judge whether two words are associated or are in the same category (Assaf et al., in press). In fact, repetitive transcranial magnetic stimulation (TMS) of the left

Downloaded from http://direct.mit.edu/jocn/article-pdf/19/5/804/1958840/jocn.2007.19.5.804.pdf by guest on 18 May 2021

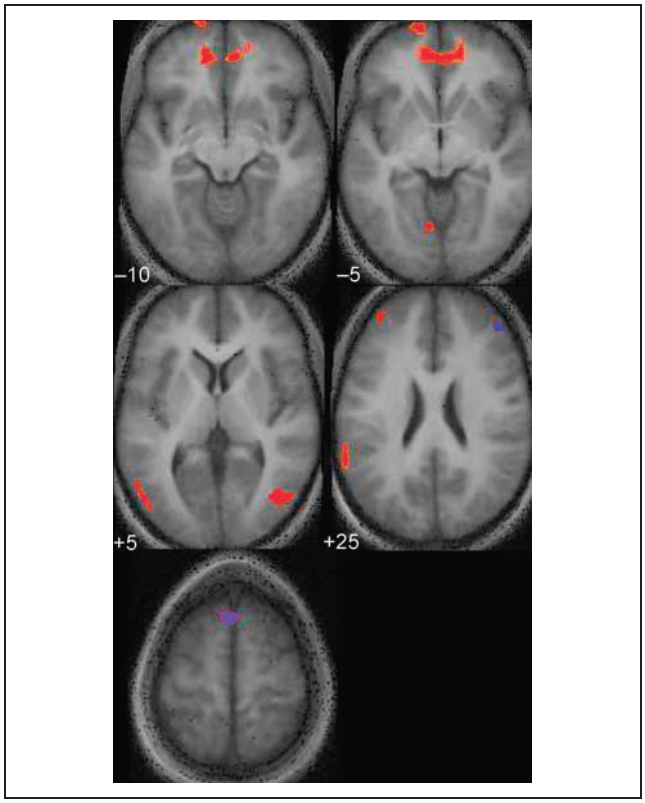

Figure 2. Main effect of morality. *F* values ($p < .05$) from the mixed-effects ANOVA are shown on a mean high-resolution anatomical T1 image of all 24 subjects in neurological convention. Regions in red were activated more in moral scenarios than in nonmoral scenarios; regions in blue were activated more in nonmoral scenarios than in moral scenarios.

DLPFC has been shown to speed up analogic reasoning without affecting accuracy, whereas the TMS of the right DLPFC had no effect (Boroojerdi et al., 2001). Given such research, it seems reasonable to surmise that subjects in the current study talked through decisions to moral questions in their heads and—perhaps, if our interpretation of our posterior STS activation is correct—tried to match aspects of difficult novel scenarios to heuristics and/or easy previously addressed scenarios for which they had answers. Nonmoral scenarios did not elicit as much dissonance and, therefore, generally utilized spatial, rather than verbally mediated, problem solving.

Some brain regions associated with moral conditions of previous fMRI morality studies were not implicated in the present study. In particular, the posterior cingulate (BA 31), an area associated with emotion (Maddock, 1999) that has been shown to be active during personal moral judgments (Greene, Nystrom, et al., 2004; Greene, Sommerville, et al., 2001) but not during simple moral judgments (Heekeren et al., 2003) or moral judgments compared to unpleasant (as opposed to neutral) nonmoral stimuli (Moll, de Oliviera-Souza, Bramati, et al., 2002), did not appear in the main effect of morality or in experimental interactions, despite the use of complex and provoking moral scenarios. These results suggest that posterior cingulate activation associated with the personal stimuli used by Greene, Nystrom, et al. (2004) and Greene, Sommerville, et al. (2001) is not elicited merely by scenario complexity or general emotional conflict. Unlike stimuli used in other studies of morality, the moral personal scenarios of Greene, Nystrom, et al. and Greene, Sommerville, et al. often referenced immediate family members or close friends. Thus, although it is possible that the posterior cingulate becomes preferentially involved in moral reasoning when the situation under consideration involves one's loved ones, this

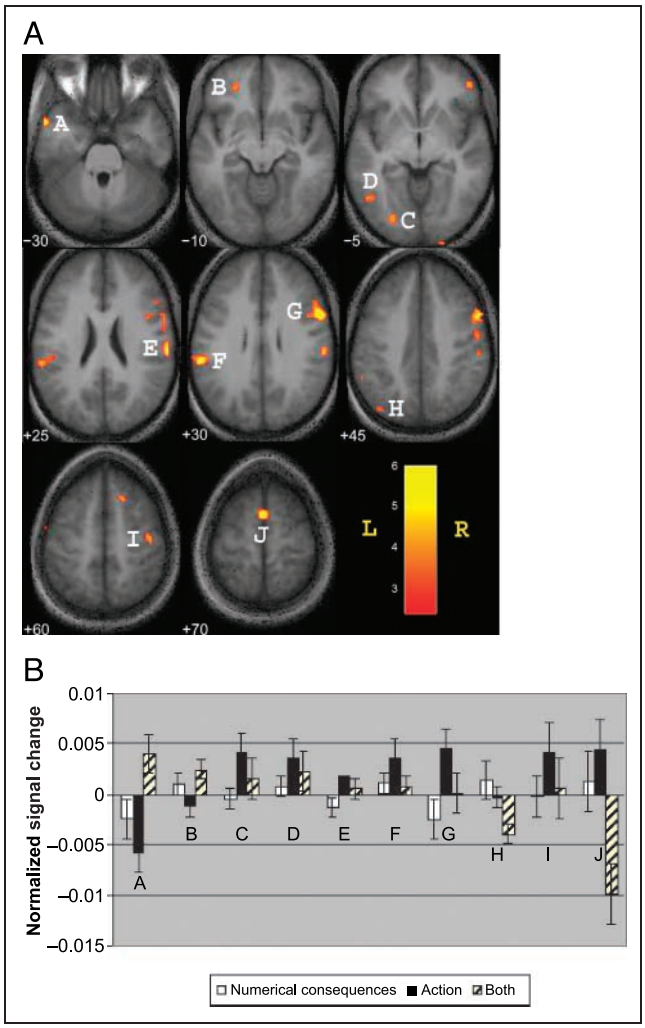

Figure 3. Morality × Type interaction. (A) *F* values ($p < .05$) from the mixed-effects ANOVA are shown on a mean high-resolution anatomical T1 image of all 24 subjects. (B) Differences in signal between moral and nonmoral scenarios of numerical consequences, action, and both conditions. A = Middle temporal gyrus (BA 38); B = lateral orbital frontal gyrus (BA 11); C = left middle occipital gyrus (BA 18); D = posterior inferior temporal sulcus (BA 37); E = right supramarginal gyrus (BA 40); F = left supramarginal gyrus (BA 40); G = middle frontal gyrus (BA 8); H = left angular gyrus (BA 40); I = right central sulcus, rostral DLPFC (BA 4); J = bilateral medial superior frontal (BA 6).

Downloaded from http://direct.mit.edu/jocn/article-pdf/18/5/803/1958840/jocn.2006.18.5.803.pdf by guest on 18 May 2021

study and past studies suggest that the posterior cingulate is not preferentially activated by all deontological or emotional moral processings.

Past studies have also implicated two regions of the paralimbic system—the amygdala (Greene, Nystrom, et al., 2004) and the temporal pole (Heekeren et al., 2003; Moll, Eslinger, et al., 2001)—in moral judgment, but these regions were not more active in moral scenarios than in nonmoral scenarios once numerical consequences, action, means, and language have been taken into account (although the temporal pole appeared in morality interactions).

Experimental Factors

Type Variable: Consequences and Doing vs. Allowing

Consequentialism claims that moral decisions should be made to maximize overall consequences (defined as number of lives in this study). The DDA, in contrast, states that it takes more to justify performing an action that results in harm than to justify refraining from an action, thereby allowing harm to occur. Consequentialism and the DDA thus prescribe that different factors should affect moral judgments. These factors did affect our subjects' answers (Table 3). In response to the question ''Would you (action appropriate to the scenario)?'', subjects said that they would perform the action resulting in a certain number of deaths over the inaction resulting in the same number of deaths only 9% of the time in moral/action scenarios. In contrast, when inaction resulted in worse overall consequences than action in moral/both scenarios, the percentage of responses choosing action over the inaction rose to 31%. When both presented options required action in moral/numerical consequence scenarios, subjects chose the best consequences 49% of the time. Furthermore, these percentages represent the number of indicated choices, irrespective of considerations of intentional harm or colorful language that also affect subject choice. Thus, subjects chose not to act in most moral scenarios, as the DDA would oblige, but they were also more likely to forego their preference for inaction if action resulted in better consequences, in accordance with Consequentialism. As reflected by their answers to the question ''Is it wrong to (action appropriate to the scenario)?'', subjects' decisions about which act (or failure to act) was wrong in a moral situation followed a pattern similar to that of their decisions about what they would do, but the pattern was no more pronounced in moral scenarios than in nonmoral scenarios, so the pattern was significant in the main effect of Type, but not in the Morality × Type interaction.

The distinctions made by the DDA and Consequentialism also affected subjects' response times (see Table 3). Most notably, if reaction time is interpreted as deliberation time (Greene, Nytsrom, et al., 2004), it can be inferred that it required less deliberation for subjects to answer ''Would you do it?'' when both alternatives within a scenario had the same consequences than when the alternatives' consequences were different.

Distinct patterns of brain activity (Figure 3) were associated with each of the three levels of moral/type scenarios—consequences, action, and both—consistent with subjects' varying responses and response times and partially consistent with our fourth hypothesis. Imaging data suggest that moral consideration of the action-vs.-inaction distinction is mediated primarily by areas of the brain that are traditionally associated with cognition rather than with emotion. Moral/action scenarios preferentially activated a large region (491 voxels) of the right middle frontal gyrus, rostral DLPFC (BA 8, Figure 3G), an area implicated in conscious downregulation of negative emotion (Ochsner et al., 2004). Consideration of the action-vs.-inaction distinction also recruited other areas associated with cognitive processing, including BA 18, 37, 40, 4, and 6 (Figure 3C–F, I, and J, respectively), more than nonmoral/action scenarios and more than the overall mean activation in those regions. On the other hand, the middle temporal gyrus (BA 38, Figure 3A) and a large region (149 voxels) of the medial orbital gyrus (BA 11, Figure 3B)—two paralimbic regions involved in emotional reinforcement processing (Rolls, 2004)—were activated much less than in nonmoral/action scenarios and much less than the overall mean activation in those regions. Put simply, contrary to our hypothesis, when consequences were held constant, moral deliberation about action versus inaction invoked activity in areas dedicated to high-level cognitive processing and suppressed activity in areas associated with socioemotional processing.

In contrast to actions and omissions, the cognitive resources used to process numerical consequences in moral scenarios did not diverge much from those elicited by nonmoral/numerical consequence scenarios. Moral/numerical consequence scenarios activated the DLPFC (BA 8, Figure 3G) less than nonmoral/numerical consequence scenarios, suggesting that less working memory was required to respond to moral/numerical consequence scenarios than their nonmoral counterparts. Considering that most people find it strongly preferable to save more lives rather than fewer lives whenever possible, it would be expected that little effort and/or deliberation requiring working memory would be recruited to choose a response in moral scenarios offering options that differed only in their consequences. When lives are not at stake, considerations such as required effort or time are likely to carry more weight and to ultimately contribute to cognitive load, as presumably illustrated in patterns of brain activation associated with nonmoral/numerical consequence scenarios.

Singular patterns of brain activity emerge, however, when numerical consequences in moral scenarios conflict with the desire to refrain from action. Both the orbital frontal cortex (Figure 3B) and the middle temporal

Downloaded from http://direct.mit.edu/jocn/article-pdf/20/5/804/1958840/jocn.2008.20057.pdf by guest on 18 May 2021

gyrus (Figure 3A), two areas previously implicated in moral decision making (Heekeren et al., 2003; Moll, de Oliveira-Souza, Bramati, et al., 2002; Moll, de Oliviera-Souza, Eslinger, Bramati, et al., 2002; Moll, Eslinger, et al., 2001), were much more activated in moral/both scenarios than in nonmoral/both scenarios and in the mean activation in those regions. Furthermore, the left angular gyrus (BA 40, Figure 3H), which is a region implicated in number processing (Dehaene, Piazza, Pinel, & Cohen, 2003), and regions of the medial superior frontal gyrus, SMA area (Figure 3J) were much less activated in moral/both scenarios than in nonmoral/both scenarios or the mean activation in those regions, indicating that areas associated with cognitive processes were less engaged when moral/both scenarios were addressed. Whereas moral/action scenarios called upon explicit cognitive processes, moral/both scenarios were managed by more stimulus-driven emotional processes. These functional imaging results imply that competition between conflicting intuitions attributable to numerical consequences and the DDA exacts more emotional and more automatic processing than either intuition acting alone without incongruity. Thus, the emotional activation that Greene, Nystrom, et al. (2004) and Greene, Sommerville, et al. (2001) attributed to personal scenarios may not be due merely to the action-vs.-inaction distinction or influences of numerical consequences in isolation, but some of the observed activation may be due to the conflict of these, or similar, variables.

Means Variable: The DDE

Subjects' responses were affected by whether harm was intended as a means, as forbidden by the DDE, but less dramatically than by the action-vs.-inaction distinction emphasized by the DDA (Table 3). Overall, subjects were more likely to answer that it was wrong to cause a harm intentionally than unintentionally, and to an even greater extent in moral scenarios than in nonmoral scenarios. Similarly, in moral scenarios, subjects' answers to ''Would you (action appropriate to the scenario)?'' indicated that they were much less likely to cause harm intentionally than unintentionally. Subjects could determine their typically negative answers more quickly when an option required intentional action over unintentional action, but no more so in moral scenarios than in nonmoral scenarios.

Moral processing of intentionally causing harm was characterized by a pattern of brain activity very different from that in the moral processing of action vs. nonaction. First, the means variable affected activation in the nonmoral scenarios much more than in the moral scenarios. Second, the pattern of brain activity associated with the means variable in the moral condition resembled the activity that Greene, Nystrom, et al. (2004) and Greene, Sommerville, et al. (2001) observed in personal moral scenarios. The greatest activation differences between the moral/means and the moral/nonmeans conditions appeared as a decrease in activity in the right angular gyrus (BA 40, Figure 4G) and as an increase in activity in anterior regions of the STS (BA 21/22, Figure 4A) and the orbital frontal cortex (BA 11, Figure 4B). The DLPFC was not preferentially activated in moral/means scenarios, such as the footbridge case, in contrast to the action condition. In moral/nonmeans scenarios, such as the sidetrack case,

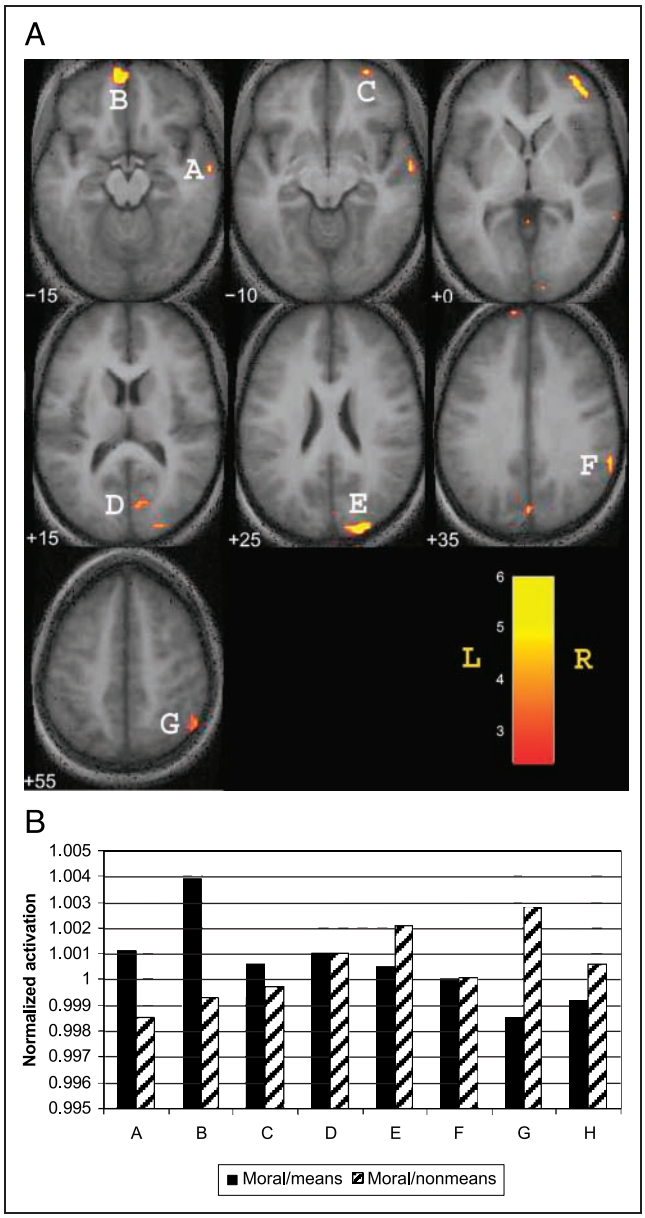

Figure 4. Morality × Means interaction. (A) F values ($p < .05$) from the mixed-effects ANOVA are shown on a mean high-resolution anatomical T1 image of all 24 subjects. (B) Normalized signal for moral/means and moral/nonmeans conditions. A = anterior STS (BA 21/22); B = medial orbital frontal (BA 11); C = right frontomarginal gyrus (BA 10); D = calcerine sulcus (BA 17); E = right superior occipital gyrus (BA 19); F = right supramarginal gyrus (BA 40); G = right angular gyrus (BA 40); H = right SMA.

Downloaded from http://direct.mit.edu/jocn/article-pdf/18/5/803/1958840/jocn.2006.18.5.803.pdf by guest on 18 May 2021

cognitive areas—including the right superior occipital gyrus (BA 19, Figure 4E), right angular gyrus (BA 40, Figure 4G), and SMA (BA 6, Figure 4H)—were preferentially activated. These results reveal that moral dilemmas requiring intentional harm to people elicit more emotional processing, similar to that elicited by the personal moral scenarios of Greene, Nystrom, et al. (2004) and Greene, Sommerville, et al. (2001), than moral dilemmas not requiring intentional harm to people. Such emotional brain activity might be responsible for subjects' tendencies to respond to means scenarios more quickly.

The greatest difference between absolute activations in the moral/means and in the moral/nonmeans scenarios (rather than between a moral means/nonmeans condition and its respective nonmoral counterpart) was the decrease in activity in the right angular gyrus (BA 40, Figure 4G, $p = .035$). The right angular gyrus is a complex heteromodal region associated with spatial attention and the ability to evaluate one's sense of agency and sense of responsibility for actions (Farrer et al., 2004; Farrer & Frith, 2002). Given this latter evidence, one can speculate that the difference in activation in this area might be because subjects felt more in control of their decisions and more accountable to their choice when they were making moral decisions that did not require intentional harm to another. Observations made of subjects after finishing the pilot behavioral study and the fMRI study are consistent with this interpretation. Subjects were usually puzzled by their violent reaction towards dilemmas such as the footbridge trolley case and frequently became frustrated in trying to justify their choices towards moral/means scenarios in informal debriefings after scanning, often resorting to statements such as ''I don't know, it's just not right!'' (see ''moral dumbfounding'', Haidt, 2001). In contrast, almost all subjects sanguinely reported the use of some form of the DDA in action/both scenarios. The tenacious intuition that something was wrong without a corresponding conscious explanation may be responsible for the underactivated angular gyrus in moral/means scenarios (Farrer et al., 2004). Subjects felt like helpless observers—not ''authors''—of their own passionate aversions to certain types of action requiring intentional harm.

Language Variable

Language manipulations were very subtle, and their behavioral effects were correspondingly modest (Table 3). Nevertheless, imaging data corroborate that they had significant effects on multiple regions of the brain after all of the other variables had been taken into account, albeit in patterns not predicted by our experimental hypotheses (Table 4).

The data are inconclusive about whether results from past studies are more consistent with scenarios described in plain language or in colorful language, and no definitive statements can be made about what the brain activity identified in the Morality × Language interaction represents, but it is clear that even very subtle language manipulations affect both moral and nonmoral decision making. The data reported here strongly suggest that future imaging studies of morality take care to standardize the amount of dramatic and descriptive language used in both moral and control conditions.

Yes/No Responses

Emotional processing did not correspond with more yes responses to the question ''Is it wrong to do it?'' or with more no responses to the question ''Would you do it?'' More people said that it was not wrong to perform the actions presented in the moral/numerical consequence scenarios than the actions in either the moral/action or the moral/both scenarios, yet comparatively, moral/numerical consequence scenarios elicited only intermediate amounts of both emotional and cognitive processing. Furthermore, more emotional processing was detected in moral/both scenarios than in moral/action scenarios, yet more people responded that they would perform the action offered in moral/both scenarios than in moral/action scenarios. In confirmation, multiple iterations of statistical analyses were performed to investigate possible correlations between brain activity and subject response choice. No statistically significant relationships were found (similarly to Moll, Oliviera-Souza, Eslinger, Bramati, et al., 2002). Thus, it is likely that the regions identified in Morality × Means and Morality × Type interactions generally reflect only how the scenarios were processed, not what subjects' final judgments or decisions would be.

Conclusion

This study has found that the moral factors that we differentiate in our ethical and political lives—consequences, action, and intention—are also differentiated by the brain processes required to make them. Our data are consistent with the possibility that separate heuristics and brain systems underlie use of the doctrines of Consequentialism, Doing and Allowing, and Double Effect. In contrast to the speculations of Greene, Nystrom, et al. (2004) and Greene, Sommerville, et al. (2001), our data suggest that some deontological responses (such as the DDA-implied intuition to refrain from action) can be mediated by reason (as suggested by Kant and, more recently, by Hauser), whereas other deontological responses (such as the DDE-implied intuition to refrain from intentional harm) can be mediated by emotion (as suggested by Hume and, more recently, by Haidt). Individuals will utilize varying combinations of cognitive and emotive facilities to address moral challenges, but, overall, certain types of moral scenarios are likely to be

Downloaded from http://direct.mit.edu/jocn/article-pdf/20/2/203/1958540/jocn.2008.20057.pdf by guest on 18 May 2021

processed in characteristic ways. Cognitive processing was associated with moral scenarios requiring a choice between a harmful action to lessen an existing threat and an inaction. The emotional patterns of brain activation isolated by Greene, Nystrom, et al. (2004) and Greene, Sommerville, et al. (2001) were primarily associated with dilemmas using people as a means to save others. We demonstrated that emotional systems also become enlisted when the intuition to avoid the creation of a threat and the intuition to minimize harm are aroused simultaneously. Still, emotional activation in the paralimbic system is not associated with an increased frequency of judgments that something is morally wrong. Disparate combinations of systems used to resolve moral dilemmas we have identified represent differences only in processing, not differences in judgments or actions. Further research is needed to elucidate the regions of the brain that are truly responsible for the choice to answer ''Yes, it is wrong'' compared to ''No, it is not wrong.''

Our data highlight that morality is not represented in one place in the brain, but instead is mediated by multiple networks. Some of these neural networks are modulated by the language in which stimuli are presented; thus, future studies on morality should control for the level of emotional and/or descriptive details in their stimuli accordingly. Likewise, careful attention needs to be given to the nature of nonmoral control conditions. Furthermore, most published fMRI studies, including this one, are designed to map out regions of the brain that are active during specific types of moral decisions. After this important initial groundwork is laid, a next step for moral researchers is to use network modeling methods to delineate how the regions of the brain identified in these first fMRI studies cooperate and interact.

We want to emphasize that this study was designed to characterize the neural processes involved in responding to moral dilemmas, not to attempt to find answers to how moral dilemmas should be processed or resolved. Studying conscious moral deliberation helps identify factors that are also likely to play a role in simple, but important, moral judgments and decisions. The doctrines operationalized in this study often facilitate our views on topics ranging from social etiquette to controversial issues of abortion or euthanasia. Understanding how these heuristics are neurologically represented may help explain why moral convictions vary so greatly, which, in turn, may influence the decisions we make as individuals and the way we understand and judge other cultures.

Acknowledgments

We thank George Wolford, Roger Norlund, and Tammy Laroche for their helpful contributions to this project.

Reprint requests should be sent to Scott T. Grafton, Center for Cognitive Neuroscience, 6162 Moore Hall, Dartmouth College, Hanover, NH 03755, or via e-mail: Scott.T.Grafton@Dartmouth.edu.

The data reported in this experiment have been deposited with the fMRI Data Center (www.fmridc.org). The accession number is 2-2006-1211A.

REFERENCES

Assaf, M., Calhoun, V. D., Kuzu, C. H., Kraut, M. A., Rivkin, P. R., Hart, J., Jr., & Pearlson, G. D. (in press). Neural correlates of the object recall process in semantic memory. *Psychiatry Research: Neuroimaging.*

Baddeley, A. (1986). *Working memory.* New York: Clarendon Press/Oxford University Press.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41,* 809–823.

Boroojerdi, B., Phipps, M., Kopylev, L., Wharton, C. M., Cohen, L. G., & Grafman, J. (2001). Enhancing analogic reasoning with rTMS over the left prefrontal cortex. *Neurology, 56,* 526–528.

Braver, T. S., & Bongiolatti, S. R. (2002). The role of frontopolar cortex in subgoal processing during working memory. *Neuroimage, 15,* 523–536.

Caplan, R., & Dapretto, M. (2001). Making sense during conversation: An fMRI study. *NeuroReport: For Rapid Communication of Neuroscience Research, 12,* 3625–3632.

Chee, M. W. L., O'Craven, K. M., Bergida, R., Rosen, B. R., & Savoy, R. L. (1999). Auditory and visual word processing studied with fMRI. *Human Brain Mapping, 7,* 15–28.

Damasio, A. R. (1994). *Descarte's error: Emotion, rationality and the human brain.* New York: Putnam.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20,* 487–506.

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Jr., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition, 92,* 145–177.

Eslinger, P. J., Flaherty-Craig, C. V., & Benton, A. L. (2004). Developmental outcomes after early prefrontal cortex damage. *Brain and Cognition, 55,* 84–403.

Farrer, C., Franck, N., Frith, C. D., Decety, J., Georgieff, N., d'Amato, T., & Jeannerod, M. (2004). Neural correlates of action attribution in schizophrenia. *Psychiatry Research: Neuroimaging, 131,* 31–44.

Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs. another person as being the cause of an action: The neural correlates of the experience of agency. *Neuroimage, 15,* 596–603.

Fischer, J. M., & Ravizza, M. (Eds.) (1992). *Ethics: Problems and principles.* Fort Worth, TX: Harcourt, Brace, and Jovanovich College Publishers.

Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87,* B11–B22.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6,* 517–523.

Greene, J. D., Nystrom, L. E., Engell, A. D., & Darley, J. M. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44,* 389–400.

Greene, J. D., Sommerville, R., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105–2108.

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

Hauser, M. D. (2006). *Moral minds: The unconscious voice of right and wrong.* New York: Harper Collins.

Downloaded from http://direct.mit.edu/jocn/article-pdf/18/5/803/1958840/jocn.2006.18.5.803.pdf by guest on 18 May 2021

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

Howard-Snyder, F. (2002). Doing vs. allowing harm. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Available at http://plato.stanford.edu/entries/doing-allowing/.

Hume, D. (1888). *A treatise of human nature.* Oxford: Clarendon Press.

Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). Neural correlates of self-reflection. *Brain, 125,* 1808–1814.

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American, 246,* 160–173.

Kant, I. (1959). *Foundations of the metaphysics of morals.* Indianapolis, IN: Bobbs-Merrill.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience, 14,* 785–794.

Koechlin, E., Basso, G., Pietrini, P., Panzer, S., & Grafman, J. (1999). The role of the anterior prefrontal cortex in human cognition. *Nature, 399,* 148–151.

Kostopoulos, P., & Petrides, M. (2003). The mid-ventrolateral prefrontal cortex: Insights into its role in memory retrieval. *European Journal of Neuroscience, 17,* 1489–1497.

Maddock, R. J. (1999). The retrosplenial cortex and emotion: New insights from functional neuroimaging of the human brain. *Trends in Neurosciences, 22,* 310–316.

McIntyre, A. (2004). Doctrine of Double Effect. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Available at http://plato.stanford.edu/entries/double-effect/.

Mesulam, M. M. (2000). Behavioral neuroanatomy: Large-scale networks, association cortex, frontal syndromes, the limbic system, and hemispheric specializations. In M. M. Mesulam (Ed.), *Principles of behavioral and cognitive neurology* (2nd ed., pp. 1–120). London: Oxford University Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202.

Moll, J., Eslinger, P. J., & Oliveira-Souza, R. (2001). Frontopolar and anterior temporal cortex activation in a moral judgment task. *Arquivos de Neuro-Psiquiatria, 59,* 657–664.

Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional networks in emotional moral and nonmoral social judgments. *NeuroImage, 16,* 696–703.

Moll, J., de Oliveira-Souza, R., & Eslinger, P. J. (2003). *Morals and the human brain: A working model.* Philadelphia, PA: Lippincott Williams & Wilkins.

Moll, J., de Oliviera-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience, 22,* 2730–2736.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment.* New York: Oxford University Press.

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D. E., & Gross, J. J. (2004). For better or for worse: Neural systems supporting the cognitive down- and up-regulation of negative emotion. *Neuroimage, 23,* 483–499.

Rolls, E. T. (2004). The functions of the orbitofrontal cortex. *Brain and Cognition, 55,* 11–29.

Sinnott-Armstrong, W. (2003). Consequentialism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Available at http://plato.stanford.edu/entries/consequentialism/.

Suchan, B., Yagueez, L., Wunderlich, G., Canavan, A. G. M., Herzog, H., Tellmann, L., Homberg, V., & Seitz, R. J. (2002). Hemispheric dissociation of visual-pattern processing and visual rotation. *Behavioural Brain Research, 136,* 533–544.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain.* New York: Thieme Medical Publishers, Inc.

Downloaded from http://direct.mit.edu/jocn/article-pdf/19/5/798/1958840/jocn.2007.19.5.798.pdf by guest on 18 May 2021