doi:10.1093/scan/nsq011 SCAN (2010) 5, 404–413

# On the wrong side of the trolley track: neural correlates of relative social valuation

**Mina Cikara,[1] Rachel A. Farnsworth,[2] Lasana T. Harris,[3] and Susan T. Fiske[1]**

[1]Princeton University, Princeton NJ 08540, [2]University of Pennsylvania Law School, Philadelphia PA 19104, and [3]New York University, New York NY 10003

**Using moral dilemmas, we (i) investigate whether stereotypes motivate people to value ingroup lives over outgroup lives and (ii) examine the neurobiological correlates of relative social valuation using fMRI. Saving ingroup members, who seem warm and competent (e.g. Americans), was most morally acceptable in the context of a dilemma where one person was killed to save five people. Extreme outgroup members, who seem neither warm nor competent (e.g. homeless), were the worst off; it was most morally acceptable to sacrifice them and least acceptable to save them. Sacrificing these low-warmth, low-competence targets to save ingroup targets, specifically, activated a neural network associated with resolving complex tradeoffs: medial PFC (BA 9, extending caudally to include ACC), left lateral OFC (BA 47) and left dorsolateral PFC (BA 10). These brain regions were recruited for dilemmas that participants ultimately rated as relatively more acceptable. We propose that participants, though ambivalent, overrode general aversion to these tradeoffs when the cost of sacrificing a low-warmth, low-competence target was pitted against the benefit of saving ingroup targets. Moral decisions are not made in a vacuum; intergroup biases and stereotypes weigh heavily on neural systems implicated in moral decision making.**

**Keywords:** fMRI; social valuation; moral dilemmas; intergroup bias

People make moral tradeoffs every day, deciding whether to endorse welfare policies that help a few at the expense of the many, whether to support a war that risks soldiers and outgroup civilians for the apparent good of the ingroup, or whether to donate to charities that support the less fortunate at one's own expense. How do people decide what is preferable? Do they maximize the number of people who benefit and minimize the number who suffer, or do biased value assessments lead them to favor fewer, perhaps socially preferred lives? We suggest that people's resolutions of moral tradeoffs are driven in part by intergroup biases: ingroup favoritism plus differential devaluation of various outgroups.

In any tradeoff, people strive to protect their own social groups—their ingroups—at the expense of outgroups to which they do not belong (Tajfel and Turner, 1986; Brewer, 1999). An open question remains regarding whether people differentially value the lives of some outgroup members over other outgroup members (Cuddy *et al.*, 2007). Behavioral and neuroscience evidence suggests that people do perceive some outgroups as more human than others, and by extension, perhaps more valuable (Harris and Fiske, 2006, 2009; Haslam, 2006). In a tradeoff, people may be particularly inclined to benefit someone important (e.g. an ingroup member) at the cost of someone they perceive to be 'worth' less.

Received 21 June 2009; Accepted 13 January 2010
Advance Access publication 11 February 2010

We wish to thank the Russell Sage Foundation, the Princeton Neuroscience Institute, Princeton University's Roundtable Fund and the Princeton University Psychology Department for their generous support of this research project. We would like to thank Joshua Greene, Daniel L. Ames, Adam Moore and Amitai Shenhav for their helpful comments. This research is based on the second author's senior thesis, and parts were presented at the annual meeting of the 2008 American Psychological Association, Boston, MA.

Correspondence can be addressed to Mina Cikara, Department of Psychology, Princeton University, Princeton NJ 08540, USA. E-mail: mcikara@princeton.edu.

## INTERGROUP BIAS: PROTECT THE INGROUP

Group identity engenders ingroup favoritism, which in turn reinforces the boundaries between social categories (e.g. favoring 'us' versus 'them'; Tajfel and Turner, 1986; for review see Hewstone *et al.*, 2002). Favoritism most explicitly manifests in resource allocation; groups reserve resources for those they favor and withhold resources from those they derogate. Moral emotions by definition relate to the welfare of the ingroup (Moll *et al.*, 2003); however, moral emotions such as contempt or xenophobia may also promote social conflict (Rozin *et al.*, 1999). In overt conflict, the most extreme intergroup biases delegitimize victims: deeming them outcasts, who do not deserve protection (Bar-Tal, 1989), and morally excluding them by placing them outside the boundary of justice that applies to the ingroup (Opotow, 1995; Staub, 2001).

Evolutionary biologists and moral philosophers suggest certain dimensions matter when people make moral tradeoffs (Petrinovich *et al.*, 1993): speciesism (animal *vs* human), abhorrent political philosophy (Nazi *vs* not), inclusive fitness (kin *vs* stranger) and elitism (high- *vs* low-status person), among others. For example, participants reported that it was acceptable to kill Nazis in order to save innocent bystanders. More recently (Uhlmann, Pizarro, Tannenbaum, and Ditto, 2009), participants reported it is more morally appropriate to sacrifice 'Tyrone Payton' (named to suggest he is

African American) to save '100 members of the New York Philharmonic' (mostly white) than it is to sacrifice 'Chip Ellsworth III' (named to suggest he is white) to save '100 members of the Harlem Jazz Orchestra' (mostly black). Indeed, ingroup loyalty/betrayal may be one of the fundamental foundations of morality (Haidt and Graham, 2007); thus, group membership and associated stereotypes should factor into the processes engaged when people face moral tradeoffs involving ingroup and outgroup members.

### Stereotype content model

Often automatically activated, stereotypes may help people determine quickly whether some lives are more valuable than others (for review, see Macrae and Bodenhausen, 2000). The Stereotype Content Model (SCM; Fiske *et al.*, 2002) organizes group stereotypes along two fundamental social dimensions: competence and warmth (Fiske *et al.*, 2007). This 2 (low/high warmth) × 2 (low/high competence) mapping describes four broad stereotype categories and associated emotional responses (Figure 1). Groups high on both warmth and competence are the ingroup (e.g. Americans), and people respond to them with pride, whereas targets low on both warmth and competence (e.g. homeless) are the most extreme outgroups at the bottom of the social hierarchy and elicit emotions such as disgust (Harris and Fiske, 2006). Targets falling in the mixed quadrants elicit ambivalent emotions; envy is reserved for targets perceived as high competence but low warmth (e.g. professionals), and pity is elicited by targets perceived as low in competence and high in warmth (e.g. elderly) (Harris and Fiske, 2006). Our central questions are: Does the ingroup have an advantage over all outgroups in tradeoffs regarding the most valuable resource of all—life? Furthermore, are some outgroups preferentially (de)valued?

Utilizing the SCM allows us to go beyond the scope of extant research to examine relative valuations of socially relevant outgroups (i.e., many of whom are commonly targeted as recipients of welfare policy: disabled, homeless, elderly, drug addicted). In contrast to the previous studies, this work emphasizes that moral tradeoffs may be more complicated than simply benefitting the ingroup at the expense of the outgroup because not all outgroups are equivalent. Specifically, systematic principles categorize the groups based on their perceived warmth and competence. Because these dimensions are good predictors of emotional and behavioral responses to a variety of social groups (Fiske *et al.*, 2007; Cuddy *et al.*, 2008), it stands to reason that warmth and competence will also be good predictors of social valuation. Finally, examining the effects of the dimensions, not the groups themselves, allows predictions about the value of any social group based solely on the stereotype content.

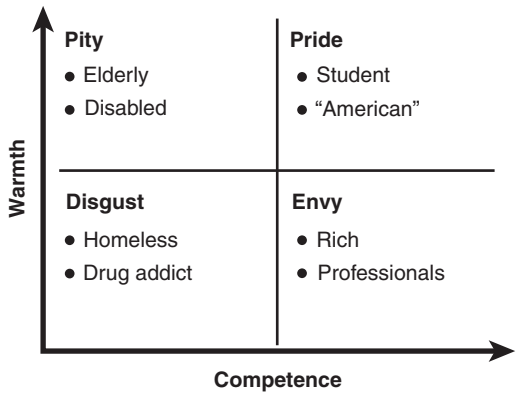


**Fig. 1** The stereotype content model warmth by competence space, stereotyped group exemplars and associated emotions. Source data from Fiske *et al.* (2002).

### Relative moral valuations

Moral philosophy has long examined the principles that people apply when forced to choose between aversive alternatives, particularly life and death. Often termed 'moral dilemmas', resolution of these scenarios recruits people's inherent sense of right and wrong. The footbridge version of the trolley dilemma (Foot, 1978; Thomson, 1986) is one such scenario: An empty runaway streetcar speeds down the tracks toward five people. Joe, from an overpass, sees this accident unfolding. If Joe chooses, he can shove a bystander off the overpass to block the streetcar, saving the five people.[1] How morally acceptable is it for Joe to push the bystander off the overpass? Here we ask: does moral acceptability depend on the stereotypes associated with the person being sacrificed and the people being saved?

The footbridge dilemma constitutes a high-conflict scenario because there is no consensus as to whether using the person on the overpass as a means to saving the other five people is acceptable (e.g. Cushman *et al.*, 2006; Koenigs *et al.*, 2007). Utilitarian reasoning aims to maximize the number of people saved in this scenario (Mill, 1861/1998), whereas deontological reasoning (Kant, 1785/1959) posits that the rights of the individual on the overpass outweigh utilitarian considerations. By at least one account, 88% of people report that pushing the person is unacceptable, suggesting that the default response in this particular scenario is driven by deontological reasoning for a large majority of respondents (Hauser *et al.*, 2007). The current study examines whether participants' ratings of acceptability shift under manipulations of the parties sacrificed and saved.

This scenario is useful for present purposes because it forces people to weigh alternatives that may reveal spontaneous biases, which are otherwise difficult for experimenters to detect and for participants to report. In other words, participants may apply different rules to determine moral acceptability, depending on who is sacrificed and who is saved. Moreover, the footbridge dilemma creates a situation in which stereotypes associated with the sacrificed and saved targets constitute the only information available for

1 Traditionally it is a runaway trolley, not a streetcar, and a footbridge instead of an overpass, but we want to suggest this event is taking place in a city setting where the juxtaposition of various social groups is highly likely.

participants' consideration. Participants may either rely on stereotype content to guide their moral judgments, or treat all targets equivalently. Hence, people's ratings of the acceptability of Joe's actions should reveal if and when they value certain lives above others.

Our aim is not to interrogate the processes underlying different kinds of moral judgments; other researchers have given that and related questions an expert treatment beyond the scope of the current investigation (e.g. Greene *et al.*, 2001, 2004, 2007; Haidt, 2001; Mendez *et al.*, 2005; Cushman *et al.*, 2006; Borg *et al.*, 2006; Valdesolo and DeSteno, 2006; Koenigs *et al.*, 2007; Young *et al.*, 2007). Rather, we employ moral dilemmas as a vehicle to examine differential value of social targets. In line with the recommendations of McGuire *et al.* (2009), we hold the moral dilemma constant and use stimuli differing only in the variable of interest (i.e. stereotype content).

## HYPOTHESES

The current investigation examines whether stereotype content modulates evaluations of moral dilemmas. Countering the demand characteristics inherent in these questions, we supplement the behavioral measures with neuroimaging data, using well-established patterns of neural activation associated with computing cost–benefit analyses and attitudinal complexity.

### Self-report ratings

Alternative hypotheses predict how people decide which moral tradeoffs are morally acceptable and which are not. According to a 'warmth primacy hypothesis', warmth (friendliness, trustworthiness), or its lack, is the more diagnostic of impression formation's two main dimensions (Wojciszke *et al.*, 1998); people may find it more acceptable to sacrifice low-warmth targets and to save high-warmth targets because that dimension tracks friends *vs* foes. In contrast, an 'economic valuation hypothesis' suggests that people will engage a cost–benefit analysis of lifetime output potential for sacrificed and saved targets (Lenton, 2002), finding it more acceptable to sacrifice low-competence targets and save high-competence targets.

The 'SCM hypothesis' posits that warmth and competence should interact to predict behavioral and neural responses to sacrificing and saving targets from each quadrant. Saving high-warmth, high-competence targets should be the most morally acceptable because these people represent the ingroup, unlike the other quadrants that contain only outgroups (Cuddy *et al.*, 2008; Fiske *et al.*, 2002). Other combinations of warmth and competence elicit unique prejudices (and emotions) toward outgroups, which may determine patterns of social valuation: high-warmth, low-competence targets elicit paternalistic prejudice (pity)—targets have subordinate status but receive pitying positivity; low-warmth, high-competence targets elicit envious prejudice (envy)—they seem admittedly able but are disliked (Cuddy *et al.*, 2008; Fiske *et al.*, 2002; Harris and Fiske, 2009). Low-warmth, low-competence targets elicit disgust, which is a qualitatively different response than the other two outgroup quadrants elicit (hence the interaction hypothesis; Harris and Fiske, 2006). Disgust is not a combination of envy and pity or a lack of pride; more important it is not even a uniquely social emotion, as it can target inanimate objects (Moll *et al.*, 2005). Finally, low-warmth, low-competence targets are demonstrably dehumanized; targets in the other three cells are not (Harris and Fiske, 2009). Therefore, we hypothesize that (i) ingroups will be least likely to be sacrificed and most likely to be saved and (ii) not all outgroups will be valued equivalently, because low-warmth, low-competence targets uniquely elicit disgust; they will be most likely to be sacrificed and least likely to be saved.

### Neural activity

All of the current dilemmas are high-conflict tradeoff scenarios; thus the comparison among the conditions should not replicate results associated with processing different kinds of moral dilemmas *per se* (e.g. Greene *et al.*, 2001, 2004; Borg *et al.*, 2006). Instead our predictions draw on neuroeconomic and social neuroscience literature examining the neural networks associated with cost-benefit analysis and attitudinal complexity. Specifically, each dilemma contains both a negative element (i.e. Joe pushes one person off an overpass) and a positive element (i.e. Joe saves five people). In deciding the moral acceptability of Joe's actions, participants will have to integrate the associated costs and benefits. We predict that the value associated with the corresponding costs and benefits will change as the targets' stereotype content changes, such that relatively more acceptable tradeoffs (e.g. sacrificing a homeless person targets to save 5 students) will ironically pose the most complex dilemma. In other words, the computation is simple when a valued person is being pushed off the overpass: unacceptable. In contrast, when a devalued person is being sacrificed, the dilemma becomes more complicated because a generally unacceptable tradeoff (killing someone, even to save others) may suddenly seem more acceptable. This more acceptable tradeoff should preferentially engage a frontal network (including OFC, dorsal and medial PFC) implicated in integrating cost–benefit information and resolving complex decisions (Montague *et al.*, 2006; Wallis, 2007; Rangel *et al.*, 2008). Again, most of the tradeoffs should be seen as somewhat unacceptable—people are generally averse to murdering others as means to an end (Hauser *et al.*, 2007)—but as the tradeoff becomes more ''worthwhile,'' participants will have to resolve the increased ambivalence associated with Joe's actions (e.g. Moll and Oliveira-Souza, 2007). Cunningham, Raye, and Johnson (2004) reported that anterior cingulate, frontal pole, and lateral OFC activity correlated positively with participants' subjective ratings of ambivalence on a variety of ''hot'' social issues (e.g. abortion, murder, welfare). In

summary, we anticipate greater activity in dorsolateral and medial PFC, lateral OFC, and anterior cingulate in response to tradeoffs that participants rate as being relatively more acceptable, as compared to those they rate as less acceptable, precisely because those are the scenarios in which participants will entertain the idea, however reprehensible, that the positives might outweigh the negatives.

## METHODS

### Participants

Participants were 18 Princeton students (12 female). All were right handed, native English speakers without reported psychiatric or neurological problems ($M_{age} = 20.7$). All participants were American citizens; 15 self-identified as 'white', 3 self-identified as 'black/African American'. On a scale of 1 (very liberal) to 9 (very conservative), participants reported an average of 4.6 for social issues and 4.4 for economic concerns. Five reported prior experience with moral dilemmas, though none were familiar with the specific dilemma presented. Behavioral responses from five participants were lost to technical problems with the response-recording box, so analyses of self-reports are based on data from 13 participants.

### Stimuli

Several prototypes fit each SCM category (Fiske *et al.*, 2002). We chose two prototypes from each quadrant, aiming to minimize confounds (i.e. race, religion, gender; see Figure 1). Ultimately, 8 stereotyped groups, 16 images per group, yielded 128 images in total. Previous findings validated that these images elicited their respective emotions as SCM predicts: where chance responding would be 25% in this forced-choice format, participants report feeling pride in response to high-warmth, high-competence targets 70% of the time, envy in response to low-warmth, high-competence targets 52%, pity in response to high-warmth, low-competence targets 83% and disgust in response to low-warmth, low-competence targets 64% (Harris and Fiske, 2006). Images were neither labeled (i.e. by stereotype or quadrant), nor did we reference targets' group membership at any point before the debriefing.[2]

Participants reported whether it was acceptable for Joe to push one person off an overpass to save five people, in 128 dilemmas that varied group members from the SCM quadrants in the positions of 'sacrificed' and 'saved' targets.[3]

### Procedure

Prior to scanning, participants familiarized themselves with a schematic of the footbridge version of the trolley dilemma. We told participants to imagine that Joe always makes the sacrifice, holding the probability of Joe's sacrifice constant across the scenarios.[4] Participants then viewed practice stimuli on a computer screen in the waiting area: they first saw a picture of one person and then a collage of five people. They were instructed to think of the person in the 1-person picture as the person sacrificed by Joe's action and the group of people in the 5-person picture as those saved. The practice dilemmas used only sacrificed and saved targets that were matched for SCM group (e.g. Joe sacrifices one rich person to save a group of five rich people), to minimize participants' suspicions regarding the study's true purpose.

Once in the scanner, participants viewed these images via an angled mirror attached to the radio-frequency coil above their eyes. Dilemmas appeared in a series of 8 blocks of 16 trials each, in a periodic event-related design (see Figure 2). After the prompt, 'To what extent was this action morally acceptable?' participants responded on a 4-point Likert scale (1 = not at all; 4 = very). The total size of the 1-person picture was equivalent to the collaged 5-person picture. For stimulus presentation, E-prime (version 1.2, Psychology Software Tools, Inc.: http://www.pstnet.com) randomly selected a 1-person image for the sacrificed target and a 5-person collage image for the saved targets. Participants responded using a fiber-optic touchpad (Current Designs Inc.: http://www.curdes.com/response) held in their right hands.

After the scan, participants provided demographic information and completed a post-dilemma questionnaire (to ensure they understood the dilemma and to probe for suspicion regarding the SCM exemplars). Participants were then debriefed and compensated.

### fMRI acquisition

Each session began with a high-resolution T1-weighted anatomical image (T1-MPRAGE, $0.5 \times 0.5 \times 1.0$ mm) for registering activity to each participant's anatomy and for spatially normalizing data across participants. Echo-planar images were acquired using a 3.0 T Siemens Allegra head-dedicated scanner (Siemens, Erlangen, Germany) with a standard 'bird-cage' head coil (TR = 2000 ms, TE = 30 ms, 196 mm FOV, matrix size = $64 \times 64$). Near whole-brain coverage was achieved with 32 interleaved 3-mm axial slices.

### fMRI preprocessing and data analysis

We preprocessed and analyzed the imaging data with Analysis of Functional Neuro-Images (AFNI; Cox, 1996)

2 One strength of using pictures of the targets instead of verbal descriptions of the scenario is that we control for confounds associated with word count and use of more or less colorful language. Granted, images of homeless people may be more arousing than images of business people, for example, but given that the focus of the current investigation is examining relative valuation of worth, we believe that images are more ecologically valid than category labels (e.g. 'elderly'). Ecological validity is especially relevant for moral cognition studies, because morality depends strongly on situational and cultural context (Casebeer, 2003; Moll et al., 2005).

3 Note that the group of five people saved comprised five members of the same stereotyped group. For example, Joe could sacrifice a homeless person (a low-low target), to save five rich people (low-warmth, high-competence targets).

4 We hold the probability of Joe's actions constant because expected moral value computations are sensitive not only to the number of lives lost and saved, but also to the likelihood that lives will be lost and saved (Shenhav & Greene, 2008). Probability is set at 1 in order to simplify the scenario and to keep participants focused on the relative values of the sacrificed and saved lives.

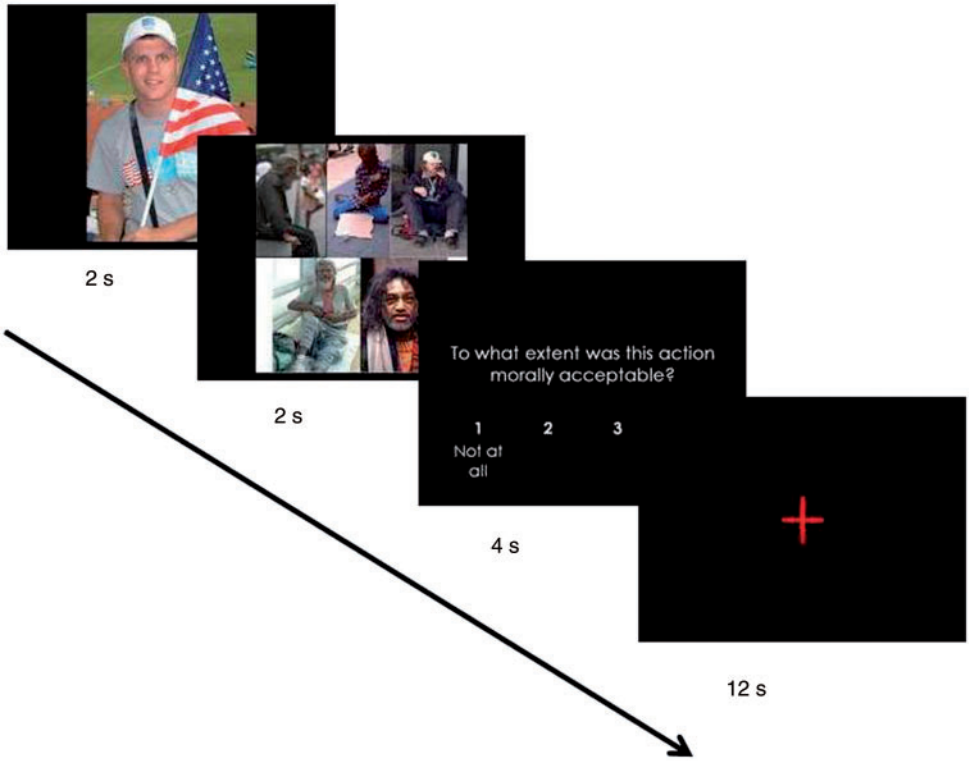


**Fig. 2** An example of a stimulus presentation block. Each dilemma comprised the following sequence: a picture of the 1 person sacrificed (2 s), a collaged picture of the five people saved (2 s), a response prompt regarding the acceptability of Joe's actions (4 s), followed by an interstimulus interval of 12 s, during which participants passively viewed a fixation cross in the center of the screen, allowing the hemodynamic response to return to baseline after each trial. The targets in each dilemma were randomly selected (see Methods section).

using standard preprocessing procedures. Participant motion was corrected using a six-parameter 3D motion-correction algorithm following slice scan-time correction. We then subjected the data to spatial smoothing with an 8-mm full width at half minimum Gaussian kernel. Finally, the signal was normalized to percent signal change from the mean.

Task-related activity was measured using a window of 4 s surrounding (2 prior to, and 2 following) the onset of the acceptability rating prompt (see Figure 2). For statistical analysis, each stimulus time series was convolved with a hemodynamic response function to create a unique regressor for each of the 16 combinations. Regressors of noninterest were also included in the multiple regression model to factor out variance associated with mean, linear and quadratic trends in each run as well as participant head motion. We used the nine-parameter landmark method of Talairach and Tournoux (1988) to normalize spatially the activation maps across participants.

Whole-brain exploratory analyses were performed with a voxelwise significance threshold of $P < 0.001$. We used the AlphaSim program included in AFNI to correct for multiple comparisons. A Monte Carlo simulation determined a minimum cluster size of 34 voxels to achieve corrected $P < 0.05$ for whole-brain contrasts, with a voxelwise threshold of $P < 0.0001$.

### Whole-brain contrasts

The contrasts between parameter estimates for different events within each participant were submitted to a group analysis of variance (ANOVA), treating between-participants variability as a random effect. Statistical parametric maps were derived from the resulting *t*-values associated with each voxel.

In an exploratory analysis, AFNI's 3dANOVA3 conducted a 2 (high/low warmth sacrificed) × 2 (high/low competence sacrificed) × 2 (high/low warmth saved) × 2 (high/low competence saved) × 18 (participants) mixed-effects ANOVA. This allowed us to examine the main effects of warmth and competence for sacrificed and saved targets, respectively. In order to address our hypotheses that ingroups would be least likely to be sacrificed and most likely to be saved but that not all outgroups would be valued equivalently, we also ran the analysis treating the SCM groups as four levels of one factor, which allowed us to specify a + 3, −1, −1, −1 contrast (for precedent, see Harris and Fiske, 2006). AFNI's 3dANOVA3 program conducted a 4 (sacrificed targets from each SCM quadrant) × 4 (saved targets from each SCM quadrant) × 18 (participants) mixed-effects ANOVA. Planned contrasts (3, −1, −1, −1) examined which voxels were more active when one quadrant exemplar was sacrificed or saved relative to the other three quadrant exemplars (e.g. which voxels were more active when Joe sacrificed a

low-warmth, low-competence target as compared to sacrificing any of the other types of targets, irrespective of the group being saved). Finally, we followed these contrasts with a single 15:1 planned contrast—comparing the sacrificing low-warmth, low-competence targets to save high-warmth, high-competence targets condition against the other 15 conditions—to examine whether a relatively more acceptable tradeoff preferentially engages the frontal network. AFNI's 3dANOVA3 program conducted a 16 (all sacrificed-saved combinations) × 18 (participants) mixed-effects ANOVA, which allowed us to specify a 15:1 contrast.

### Correlations with acceptability

We computed all correlations with acceptability scores within brain regions that were first functionally defined by the 15:1 contrast. AFNI computed the average parameter estimates for the sacrifice low–low to save high–high trials across all voxels in regions that surpassed the multiple comparisons threshold designated by AlphaSim, for each participant, individually. We then examined the correlation between average (not peak) activity in those regions and participants' respective sacrifice low–low to save high–high acceptability ratings. Note that acceptability ratings were not included in the GLM and were therefore not used to define the regions to ensure independence of the analyses (Vul *et al.*, 2009).

## RESULTS

### Behavioral data

Reliability analyses confirmed high consensus between participants' acceptability ratings for the 16 different scenarios, $\alpha = 0.93$. A 2 (high/low warmth of sacrificed target) × 2 (high/low competence of sacrificed target) × 2 (high/low warmth of saved group) × 2 (high/low competence of saved group) within-subjects ANOVA predicted participants' ratings of the moral acceptability of Joe's actions. Consistent with the economic valuation hypothesis, the main effect of competence of the sacrificed target was significant, $F(1, 12) = 6.53$, $P < 0.05$, $\eta_p^2 = 0.35$, such that it was more acceptable to sacrifice an incompetent person ($M = 2.67$, s.d. = 0.57) than a competent person ($M = 2.34$, s.d. = 0.50). The main effect of warmth of the sacrificed target, however, was not significant, $F(1, 12) = 2.52$, *ns*, $\eta_p^2 = 0.17$, nor was the interaction between warmth and competence significant, $F(1, 12) = 0.26$, *ns*, $\eta_p^2 = 0.02$.[5]

For the saved group, both main effects—warmth, $F(1, 12) = 11.32$, $P < 0.05$, $\eta_p^2 = 0.49$ and competence, $F(1, 12) = 12.82$, $P < 0.05$, $\eta_p^2 = 0.52$—were significant on moral acceptability ratings, consistent respectively with the warmth primacy and economic valuation hypotheses. In line with the SCM hypothesis, the main effects were qualified by a significant interaction, $F(1, 12) = 7.50$, $P < 0.05$, $\eta_p^2 = 0.36$; it was least acceptable to save low-warmth, low-competence targets (extreme outgroup) and most acceptable to save a group of high-warmth, high-competence targets (the ingroup) (see Figure 3). Means for all four groups differ significantly from the scale's low-endpoint, 1 [all $t(12) > 6.5$, $P < 0.05$], and from one another in paired *t*-tests [all $t(12) > 2.3$, $P < 0.05$, except high warmth, high competence *vs* low-warmth, high-competence paired, which was marginal, $t(12) = 2.10$, $P = 0.06$].[6] No higher order interactions between sacrificed and saved targets were significant.

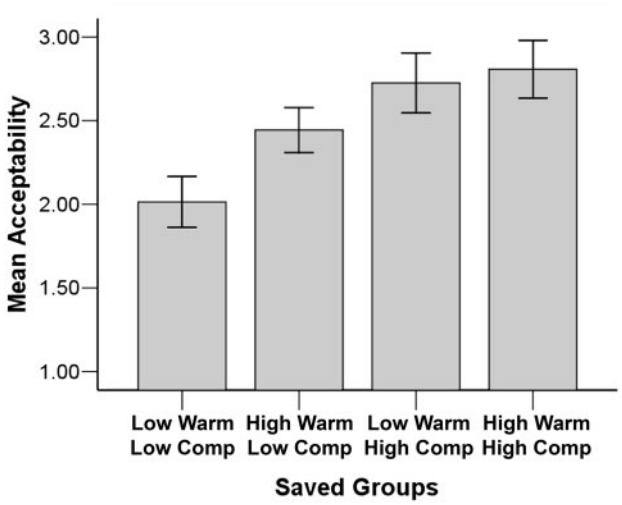


**Fig. 3** Warmth by competence interaction predicting moral acceptability of saving targets from each of the four SCM quadrants. Bars represent standard error.

Because we had an *a priori* hypothesis regarding the most acceptable tradeoff (i.e. sacrifice low warmth, low competence to save high warmth, high competence), we conducted a paired-samples *t*-test comparing that condition ($M = 2.96$) against the mean of all the other conditions ($M = 2.47$). Sacrificing low warmth, low competence to save high warmth, high competence was on average significantly more acceptable than the other conditions, $t(12) = 3.82$, $P < 0.05$.

### fMRI data

Exploratory analyses examining the main effects of sacrificing and saving high- *vs* low-warmth and high- *vs* low-competence targets are summarized in Table 1. Trials sacrificing low-competence targets, irrespective of warmth, activated a region of left middle occipital gyrus when contrasted against trials sacrificing high-competence targets. Trials saving high-competence targets activated left anterior cingulate

5 Closer inspection of the means revealed that the high warmth, low competence cell was driving an increase in acceptability of sacrificing warm targets. It was most acceptable to sacrifice low–low targets ($M = 2.73$, s.d. = 0.58), followed by high warmth, low competence targets ($M = 2.58$, s.d. = 0.60), followed by low warmth, high competence targets ($M = 2.39$, s.d. = 0.55) and least acceptable to sacrifice high–high targets ($M = 2.29$, s.d. = 0.48). Paired samples *t*-tests demonstrated that sacrificing low-low targets was more acceptable than sacrificing low warmth, high competence targets, $t(12) = -2.70$, $P < 0.05$, and high–high targets, $t(12) = -2.57$, $P < 0.05$, but not significantly more acceptable than sacrificing high warmth, low competence targets (e.g. elderly, disabled), $t(12) = -1.31$, ns.

6 One possible reason for not getting the predicted SCM interaction for the sacrificed groups but instead getting it only for the saved groups is that much intergroup discrimination is driven by benefiting the ingroup rather than harming outgroups (Mummendey, 1995).

**Table 1** Brain regions exhibiting differential activity for sacrificing and saving high- *vs* low warmth and high- *vs* low-competence targets

| Regions | Right/Left | Brodmann's area | Max *t*-score (df = 17) | Cluster size (voxels) | Talairach coordinates (*x*, *y*, *z*) |
|---|---|---|---|---|---|
| Sacrificing | | | | | |
| High warmth > Low warmth | | | | | |
| Low warmth > High warmth | | | | | |
| High comp > Low comp | | | | | |
| Low comp > High comp | | | | | |
| Middle occipital cortex | L | 19 | 4.97 | 36 | −32, −83, 20 |
| Saving | | | | | |
| High warmth > Low warmth | | | | | |
| Low warmth > High warmth | | | | | |
| High comp > Low comp | | | | | |
| Anterior cingulate | L | 32 | 5.79 | 89 | −13, 45, 7 |
| Low comp > High comp | | | | | |

Voxelwise significance threshold, $P < 0.001$, minimum cluster size 34 voxels.

**Table 2** Brain regions exhibiting differential activity for saving targets from each SCM quadrant against the average of the other three quadrants

| Regions | Right/Left | Brodmann's Area | Max *t*- score (df = 17) | Cluster size (voxels) | Talairach coordinates (*x*, y, z) |
|---|---|---|---|---|---|
| Saving | | | | | |
| High warmth, high competence | | | | | |
| Superior frontal Gyrus | L | 10 | 7.05 | 83 | −16, 55, 21 |
| Medial prefrontal Cortex | L | 9 | 5.55 | 67 | −11, 36, 43 |
| High warmth, low competence | | | | | |
| Low warmth, high competence | | | | | |
| Low warmth, low competence | | | | | |

Voxelwise significance threshold, $P < 0.001$, minimum cluster size 34 voxels.

cortex when compared against trials saving low-competence targets. None of the contrasts comparing high- and low-warmth targets when sacrificed or saved yielded significant clusters of activation.

To follow-up on the significant behavioral warmth by competence interaction for acceptability of group saved (see Figure 3), we conducted a whole-brain planned contrast (3, −1, −1, −1) examining the effect of saving targets from each SCM quadrant (against the average of the other three quadrants). Of these four, only the contrast for 'saving high-warmth, high-competence targets' yielded significant clusters of activation (see Table 2 for detailed summary of findings). Consistent with predictions, a region of mPFC was activated by a tradeoff that was rated relatively more worthwhile: saving high-warmth, high-competence targets.

The SCM also predicted an interaction between targets sacrificed and saved. Saving the ingroup by sacrificing the low-warmth, low-competence targets should differentially activate areas related to processing moral dilemmas because this combination is most acceptable. We conducted a planned 15:1 contrast comparing that condition against the remaining 15 conditions. Sacrificing the low-warmth, low-competence targets to save high-warmth, high-competence targets yielded significant clusters of activation in the following regions: medial PFC (BA 9, extending caudally to include ACC), left lateral OFC (BA 47), left dorsolateral PFC (BA 10), bilateral precuneus (BA 7) and left posterior cingulate (BA 30) (see Table 3 and Figure 4). Consistent with predictions, tradeoffs rated as more acceptable activated mPFC, DLPFC, lateral OFC and ACC. Moreover, during sacrifice of low–low to save high–high trials, average activation in OFC and DLPFC regions (identified by the 15:1 contrast) was positively correlated with acceptability of that tradeoff, $r(11) = 0.29$ and 0.31, respectively, though these relationships were not significant due to the small sample size. In other words, participants exhibiting greater activation in OFC and DLPFC in response to sacrificing low-warmth, low-competence targets to save high-warmth, high-competence targets were also those that said the tradeoff was relatively more morally acceptable.

## DISCUSSION

The stereotype content led participants to value some lives over others in moral tradeoff scenarios. Participants most endorsed saving ingroup members—Americans and students, who seem both warm and competent. Furthermore, participants did not value different outgroup members' lives equivalently. Targets belonging to extreme outgroups (i.e. low-warmth, low-competence targets) became targets of relative moral exclusion. It was most morally acceptable to sacrifice them and least acceptable to save them.

**Table 3** Brain regions exhibiting differential activity for sacrificing low warmth, low competence to save high warmth, high competence compared to all other conditions

| Regions | Right/Left | Brodmann's area | Max *t*-score (df = 17) | Cluster size (voxels) | Talairach coordinates (*x*, *y*, *z*) |
|---|---|---|---|---|---|
| Precuneus | L | 7 | 6.46 | 387 | −17, −65, 42 |
| Posterior cingulate | R/L | 29/30 | 7.35 | 357 | −6, −48, 14 |
| Medial frontal gyrus[a] | R/L | 9/32 | 6.85 | 298 | −7, 34, 28 |
| Inferior frontal gyrus | L | 47 | 5.44 | 109 | −45, 21, 3 |
| Middle frontal gyrus | L | 6 | 6.61 | 79 | −26, −3, 54 |
| Middle temporal gyrus | L | 29 | 5.75 | 56 | −57, −51, 8 |
| Middle frontal gyrus | L | 10 | 5.55 | 44 | −28, 51, 19 |

Voxelwise significance threshold, $P < 0.001$, minimum cluster size 34 voxels.
[a]Extends caudally to include anterior cingulate cortex.

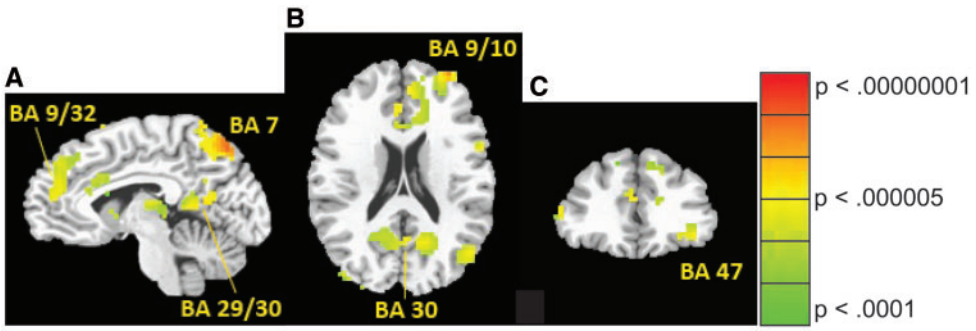


**Fig. 4** Selected brain regions (see Table 3) exhibiting significantly increased activity for sacrificing low warmth, low competence to save high warmth, high competence as compared to the other 15 conditions: mPFC (BA 9, extending caudally to include ACC), left lateral OFC (BA 47), left DLPFC (BA 10), bilateral precuneus (BA 7) and left posterior cingulate (BA 30). Statistical maps of voxelwise *t*-scores were thresholded for significance ($P < 0.001$) and cluster size (≥34 voxels). (**A**) Sagittal slice plane is $x = -6$; (**B**) axial slice plane is $z = 16$; (**C**) coronal slice plane is $y = 33$ (Talairach and Tournoux, 1988). Images (B) and (C) are reversed right to left according to radiologic convention.

Surprisingly, acceptability ratings did not depend on the warmth of the sacrificed target because it was as acceptable to sacrifice high warmth, low competence targets (elderly, disabled) as it was to sacrifice low warmth, low competence targets (homeless, drug addict). It is possible that the increased acceptability associated with sacrificing disabled and elderly people was driven by participants' lay theories regarding those targets' quality of life. While it is difficult to imagine why people would endorse shoving an individual in a wheelchair or an elderly person off an overpass, many people (both college-aged individuals as well as adults over 65) believe that health impairments, which interfere with valued life activities, constitute a "fate worse than death" (Ditto *et al.*, 1996). If participants were ambivalent about the value of high-warmth, low-competence targets' lives, they may have deemed it most acceptable for Joe to maximize the number of lives saved (i.e. preferring that Joe sacrifice one elderly or disabled person if he could save five of anyone).

On the other hand, neither warmth nor competence alone was the best predictor of moral acceptability of saving any single group of people. Rather, in concord with SCM predictions, both warmth and competence determined that it was most acceptable to save the ingroup, and least acceptable to save extreme outgroups, who are low in both warmth and competence.

## fMRI findings

Besides predicting that participants would most endorse sacrificing extreme outgroups and saving the ingroup, we predicted that these combinations would activate regions previously associated with resolving complex tradeoffs, because people would have to override their general aversion to report that some combinations were more acceptable than others (Greene *et al.*, 2001, 2004; Greene and Haidt, 2002). This network includes lateral OFC, dorsolateral and medial PFC (Wallis, 2007; Rangel *et al.*, 2008) as well as anterior cingulate cortex (Cunningham *et al.*, 2004). The only contrast that activated this network was when Joe sacrificed a low-warmth, low-competence person to save high-warmth, high-competence people (Table 3 and Figure 4). This activation dovetails nicely with the ratings data, which demonstrated that sacrificing low-warmth, low-competence people and saving high-warmth, high-competence people constituted the most acceptable classes of dilemmas.

According to recent reviews (Wallis, 2007; Rangel *et al.*, 2008), human and animal studies suggest that OFC may function to integrate multiple attributes of a decision, and

compute an associated value. Lateral PFC then utilized the value to plan behavior and medial prefrontal cortex and ACC evaluate the outcome in terms of success and required effort. By some accounts, lateral OFC contributes to information processing by inhibiting neural activity associated with unrelated or distressing information and sensations (Shimamura, 2000; Beer *et al.*, 2006). Nevertheless, studies in both humans and non-human primates have demonstrated that OFC is implicated in calculating expected value of stimuli and integrating the determined value into present and future behavior (Knutson *et al.*, 2005; Wallis, 2007). In accord with this evidence, patients with OFC damage demonstrate difficulty integrating multiple attributes in making a decision (Fellows and Farah, 2005). Note that the current data show activation in left lateral OFC, whereas others (Cunningham *et al.*, 2004) observed activation in right lateral OFC. Some evidence suggests that left lateral OFC is particularly important for the suppression of threat in decision making (Bishop *et al.*, 2004; Beer *et al.*, 2006), though it is not clear whether distinct functions engage the right *vs* left lateral OFC during inhibition (Hooker and Knight, 2006).

## General discussion

These findings suggest that even though most people say it is unacceptable to shove a person off a bridge to save five other people, utilitarian valuation of tradeoffs are demonstrably biased by the stereotype content. Specifically, 88% of people say the act is unacceptable when the targets are unidentified (Hauser *et al.*, 2007), indicating most people's default is moral aversion to the sacrifice. We reverse this pattern by manipulating the warmth and competence of the targets involved: 84% of our respondents say it is acceptable for Joe to push a low-warmth, low-competence person off a bridge to save five high-warmth, high-competence targets. We also have preliminary evidence that greater OFC and DLPFC activation was related to higher acceptability ratings of sacrificing a low-warmth, low-competence person to save high-warmth, high-competence people (though the correlations were not significant due to a small sample size). We propose that participants are actively overriding their moral aversion to using another person as a means to an end when they have the opportunity to save ingroup members by sacrificing extreme outgroup members.

What remains unclear is whether participants are actually exerting more cognitive control to override their moral aversion to sacrificing low-warmth, low-competence targets or whether they experience less moral aversion to override in the first place. Unfortunately, our design prevents claims about whether the observed activation was driven more by saving high-warmth, high-competence people (ingroup favoritism) or by sacrificing low-warmth, low-competence people in general (extreme outgroup derogation). The temporal proximity of the sacrificed and saved target images does not allow parsing of the independent effects of target sacrificed and targets saved. Future studies should either employ EEG to increase temporal resolution or provide sufficient time between the presentation of the sacrificed and saved targets to model them and behavioral responses separately.

An open question regards the other moral dilemma combinations in the study. One possibility is that exactly the same computation is occurring, only to a lesser extent, because the cost of utilizing people (other than low-warmth, low-competence outgroup targets) as a means to an end is so salient that the moral calculus is simpler: not acceptable. Recall that Greene *et al.* (2004) examine which brain regions respond more to difficult as compared to easy personal moral dilemmas and find a pattern of activation similar to our sacrifice low–low to save high–high results. Alternatively, a different, rule-based (as opposed to value-based) process may be taking place in the case of the other combinations. Well-practiced sequences (e.g. routine tasks) may be processed in posterior regions of PFC, whereas less predictable event sequences are thought to be represented in the DLPFC (Wood and Grafman, 2003).

In sum, intergroup biases and stereotypes appear to weigh heavily on neural systems implicated in moral decision making. Exactly what strategies participants used for their judgments is a complicated matter for treatment in future studies. Nevertheless, our data suggest that perceptions of warmth and competence, irrespective of the specific social groups in question, may be potent motivators in moral decision making.

**Conflict of interest.**
None declared.

## REFERENCES

Bar-Tal, D. (1989). Delegitimization: the extreme case of stereotyping and prejudice. In: Bar-Tal, D., Graumann, C., Kruglanski, A., Stroebe, W., editors. *Stereotyping and Prejudice: Changing Conceptions*. New York: Spinger-Verlag.

Beer, J.S., Knight, R.T., D'Esposito, M. (2006). Controlling the integration of emotion and cognition: the role of frontal cortex in distinguishing helpful from hurtful emotional information. *Psychological Science*, *17*, 448–53.

Bishop, S., Duncan, J., Brett, M., Lawrence, A.D. (2004). Prefrontal cortical function and anxiety: controlling attention to threat-related stimuli. *Nature Neuroscience*, *7*, 184–8.

Borg, J.S., Hynes, C., Van Horn, J., Grafton, S., Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an fMRI investigation. *Journal of Cognitive Neuroscience*, *18*, 803–17.

Brewer, M.B. (1999). The psychology of prejudice: ingroup love or outgroup hate? *Journal of Social Issues*, *55*, 429–44.

Casebeer, W.D. (2003). Moral cognition and its neural constituents. *Nature Reviews Neuroscience*, *4*, 840–6.

Cox, R.W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–73.

Cushman, F., Young, L., Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: testing three principles of harm. *Psychological Science*, *17*, 1082–9.

Cuddy, A.J.C., Fiske, S.T., Glick, P. (2008). Warmth and competence as universal dimensions of social perception: the Stereotype Content Model and the BIAS Map. In: Zanna, M.P., editor. *Advances in Experimental Social Psychology*, New York, NY: AcademicPress, pp. 61–149.

Cuddy, A.J.C., Rock, M., Norton, M.I. (2007). Aid in the aftermath of Hurricane Katrina: inferences of secondary emotions and intergroup helping. *Group Processes and Intergroup Relations*, *10*, 107–18.

Cunningham, W.A., Raye, C.L., Johnson, M.K. (2004). Implicit and explicit evaluation: FMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience*, *16*, 1717–29.

Ditto, P.H., Druley, J.A., Moore, K.A., Danks, J.H., Smucker, W.D. (1996). Fates worse than death: the role of valued life activities in health-state evaluations. *Health Psychology*, *15*, 332–43.

Fellows, L.K., Farah, M.J. (2005). Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. *Cerebral Cortex*, *15*, 58–63.

Fiske, S.T., Cuddy, A.J.C., Glick, P. (2007). Universal dimensions of social cognition: warmth, then competence. *Trends in Cognitive Sciences*, *11*, 77–83.

Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from status and competition. *Journal of Personality and Social Psychology*, *82*, 878–902.

Foot, P. (1978). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.

Greene, J., Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Neurosciences*, *6*, 517–23.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–8.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D. (2004). The neural basis of cognitive conflict in moral judgment. *Neuron*, *44*, 389–400.

Greene, J.D., Morelli, S.A., Lowenberg, K., Nystrom, L.E., Cohen, J.D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*, 1144–54.

Haidt, J. (2001). The emotional dog and its rational tail. *Psychological Review*, *108*, 814–34.

Haidt, J., Graham, J. (2007). When morality opposes justice: conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*, 98–116.

Harris, L.T., Fiske, S.T. (2006). Dehumanizing the lowest of the low: neuroimaging responses to extreme outgroups. *Psychological Science*, *17*, 847–53.

Harris, L.T., Fiske, S.T. (2009). Social neuroscience evidence for dehumanised perception. *European Review of Social Psychology*, *20*, 192–231.

Haslam, N. (2006). Dehumanization: an integrative review. *Personality and Social Psychology Review*, *10*, 252–64.

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

Hewstone, M., Rubin, M., Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, *53*, 575–604.

Hooker, C.I., Knight, R.T. (2006). Role of the orbitofrontal cortex in the inhibition of emotion. In: Zald, D.H., Rauch, S.L., editors. *The Orbitofrontal Cortex*. New York: Oxford University Press.

Kant, I. (1785/1959). *Foundation of the Metaphysics of Morals*. Indianapolis, IN: Bobbs-Merrill.

Koenigs, M., Young, L., Adolphs, R., et al. (2007). Damage to the prefontal cortex increases utilitarian moral judgements. *Nature*, *446*, 908–11.

Lenton, A.P. (2002). The price of prejudice: social categories influence monetary value of life. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *63*, 1–139.

Macrae, C.N., Bodenhausen, G.V. (2000). Social cognition: thinking categorically about others. *Annual Review of Psychology*, *51*, 93–120.

McGuire, J., Langdon, R., Coltheart, M., Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*, 577–80.

Mendez, M.F., Anderson, E., Shapira, J.S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, *18*, 193–97.

Mill, J.S. (1861/1998). In: Crisp, R., editor. *Utilitarianism*. New York: Oxford University Press.

Moll, J., de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, *11*, 319–21.

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

Moll, J., de Oliveira-Souza, R., Eslinger, P.J. (2003). Morals and the human brain: a working model. *NeuroReport: For Rapid Communication of Neuroscience Research*, *14*, 299–305.

Montague, P.R., King-Casas, B., Cohen, J.D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, *29*, 417–48.

Mummendey, A. (1995). Positive distinctiveness and social discrimination: an old couple living in divorce. *European Journal of Social Psychology*, *25*, 657–70.

Opotow, S. (1990). Moral exclusion and injustice: an introduction. *Journal of Social Issues*, *46*, 1–20.

Petrinovich, L., O'Neill, P., Jorgensen, M.J. (1993). An empirical study of moral intuitions: towards an evolutionary ethics. *Journal of Personality and Social Psychology*, *64*, 467–78.

Rangel, A., Camerer, C., Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*, 545–56.

Rozin, P., Lowery, L., Imada, S., Haidt, J. (1999). The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*, 574–86.

Shenhav, A., Greene, J.D. (2008). Representing expected moral value: outcome magnitude, probability, and expected value in the context of moral judgment. *Society for Neuroscience,* Abstract 682.12.

Shimamura, A.P. (2000). The role of prefrontal cortex in dynamic filtering. *Psychobiology*, *28*, 207–18.

Staub, E. (1989). *The Roots of Evil: The Origins of Genocide and Other Group Violence*. NY: Cambridge University Press.

Tajfel, H., Turner, J.C. (1986). The social identity theory of inter-group behavior. In: Worchel, S., Austin, L.W., editors. *Psychology of Intergroup Relations*. Chicago, IL: Nelson-Hall.

Talairach, J., Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain (M. Rayport, Trans.)*. New York: Thieme Medical Publishers.

Thomson, J.J. (1986). *Rights, Restitution, and Risk: Essays, in Moral Theory*. Cambridge, MA: Harvard University Press.

Uhlmann, E.L., Pizarro, D.A., Tannenbaum, D., Ditto, P.H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, *4*, 479–91.

Valdesolo, P., Desteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*, 476–77.

Vul, E., Harris, C., Winkielman, P., Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives in Psychological Science*, *4*, 274–90.

Wallis, J.D. (2007). Orbitofrontal cortex and its contribution to decision-making. *Annual Review of Neuroscience*, *30*, 31–56.

Wojciszke, B., Bazinska, R., Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*, 1251–63.

Wood, J.N., Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. *Nature Reviews Neuroscience*, *4*, 139–47.

Young, L., Cushman, F., Hauser, M., Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 8235–40.