Behavioral/Cognitive

Emotional and Utilitarian Appraisals of Moral Dilemmas Are Encoded in Separate Areas and Integrated in Ventromedial Prefrontal Cortex

Cendri A. Hutcherson,[1,3]* Leila Montaser-Kouhsari,[1]* James Woodward,[4] and Antonio Rangel[1,2]

[1]Division of Humanities and Social Sciences and [2]Computational and Neural Systems, California Institute of Technology, Pasadena, California 91125, [3]Department of Psychology, University of Toronto Scarborough, Toronto, Ontario M1C 1A4, Canada, and [4]Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Moral judgment often requires making difficult tradeoffs (e.g., is it appropriate to torture to save the lives of innocents at risk?). Previous research suggests that both emotional appraisals and more deliberative utilitarian appraisals influence such judgments and that these appraisals often conflict. However, it is unclear how these different types of appraisals are represented in the brain, or how they are integrated into an overall moral judgment. We addressed these questions using an fMRI paradigm in which human subjects provide separate emotional and utilitarian appraisals for different potential actions, and then make difficult moral judgments constructed from combinations of these actions. We found that anterior cingulate, insula, and superior temporal gyrus correlated with emotional appraisals, whereas temporoparietal junction and dorsomedial prefrontal cortex correlated with utilitarian appraisals. Overall moral value judgments were represented in an anterior portion of the ventromedial prefrontal cortex. Critically, the pattern of responses and functional interactions between these three sets of regions are consistent with a model in which emotional and utilitarian appraisals are computed independently and in parallel, and passed to the ventromedial prefrontal cortex where they are integrated into an overall moral value judgment.

Key words: **decision-making; emotion and cognition; medial prefrontal cortex; moral judgment; temporoparietal junction**

Significance Statement

Popular accounts of moral judgment often describe it as a battle for control between two systems, one intuitive and emotional, the other rational and utilitarian, engaged in winner-take-all inhibitory competition. Using a novel fMRI paradigm, we identified distinct neural signatures of emotional and utilitarian appraisals and used them to test different models of how they compete for the control of moral behavior. Importantly, we find little support for competitive inhibition accounts. Instead, moral judgments resembled the architecture of simple economic choices: distinct regions represented emotional and utilitarian appraisals independently and passed this information to the ventromedial prefrontal cortex for integration into an overall moral value signal.

Introduction

Many of the most consequential choices we face involve thorny questions about the appropriateness of moral tradeoffs. Is torturing a terrorist acceptable if it gains information that saves lives? How many lives must be saved to justify such an act? The profound consequences of these moral choices has inspired great interest in understanding their neural, computational, and psychological foundations. Most of this work takes the view that moral judgment involves multiple evaluative processes operating at different time-scales and levels of complexity (Greene and Haidt, 2002; Moll et al., 2005). Yet how do these processes interact? How and where are conflicts resolved? Do some processes operate more quickly or automatically than others; and if so, how are outputs from these processes combined into overall moral judgments?

Most approaches to these questions have focused on competition between two distinct systems: (1) a fast, intuitive, and largely emotion-driven system sensitive to specific features of a

Received Aug. 14, 2014; revised Aug. 5, 2015; accepted Aug. 9, 2015.

Author contributions: C.A.H., L.M.-K., J.W., and A.R. designed research; C.A.H. and L.M.-K. performed research; C.A.H. contributed unpublished reagents/analytic tools; C.A.H. and L.M.-K. analyzed data; C.A.H., L.M.-K., J.W., and A.R. wrote the paper.

This work was supported by an National Institute of Mental Health Conte Center Grant.

The authors declare no competing financial interests.

*C.A.H. and L.M.-K. contributed equally to this work.

Correspondence should be addressed to Dr. Cendri A. Hutcherson, Department of Psychology, University of Toronto Scarborough, 1265 Military Trail, Toronto, ON M1C 1A4, Canada. E-mail: chutcher@hss.caltech.edu.

DOI:10.1523/JNEUROSCI.3402-14.2015

 0270-6474/15/3512593-14$15.00/0

situation; and (2) a slower deliberative system that reasons logically about context-specific, utilitarian consequences (Greene et al., 2001, 2004; Haidt, 2001; Greene, 2005; Cushman, 2013). Although some evidence supports the distinction between these systems (Greene et al., 2008; Conway and Gawronski, 2013), their neurocomputational basis and their interactions are not fully understood. For example, are emotional appraisals the primary drivers of our moral sense, with utilitarian responses limited to *post hoc* justifications (Haidt, 2001; Wheatley and Haidt, 2005)? Do the appraisal systems compete through mutual inhibition, with the winner taking control of behavior (Greene et al., 2004; Cushman, 2013)? Or do moral judgments resemble simple economic choices, such that emotional and utilitarian appraisals are computed independently in different areas and passed to areas that integrate them into an overall value judgment (Fehr and Rangel, 2011; Lim et al., 2013; Rangel and Clithero, 2014)?

We addressed these questions using an fMRI paradigm in which human subjects provide separate emotional and utilitarian appraisals for different potential actions, and then make difficult moral judgments constructed from combinations of these actions. We build on recent work showing that the ventromedial prefrontal cortex (vmPFC) correlates with the moral value assigned to actions that vary in the distribution of lives saved (Shenhav and Greene, 2010) and that it might integrate inputs from the amygdala during difficult moral judgments (Shenhav and Greene, 2014). However, the study goes beyond these papers by identifying the separate systems involved in computing emotional appraisals, utilitarian appraisals, and overall moral value judgments. Our results support a model of moral judgment in which dissociable neural systems compute emotional and utilitarian appraisals independently and in parallel, and then pass this information to be integrated into an overall moral value in an anterior region of the vmPFC, which has been widely associated with the computation of value in economic decision-making (Bartra et al., 2013; Clithero and Rangel, 2014).

Materials and Methods

Subjects. Twenty-eight healthy individuals with normal or corrected-to-normal vision (12 female; 27 right-handed; mean age = 27.71 years; range = 19–38 years) participated in the study. We excluded data for 2 additional participants because of excessive head motion during scanning. Caltech's Institutional Review Board approved all procedures. Participants gave informed consent at the beginning of the study and were paid $40 for participating.

Study overview. At the beginning of the study, subjects received a brief overview of the three tasks they would perform in the scanner: (1) an emotional appraisal task, (2) a utilitarian appraisal task, and (3) an overall moral judgment task. The appraisal tasks were designed to identify regions specialized, respectively, in computing emotional or utilitarian appraisals for the stimuli. The overall judgment task allowed us to assess how these separate appraisals were integrated into an overall moral value for the stimuli.

The experiment consisted of one run each of the appraisal tasks and three runs of the moral judgment task. Presentation order for the emotional and utilitarian appraisal tasks was counterbalanced across subjects. The judgment task always occurred afterward. Participants' heart rate, respiration, and eye movements were recorded for all functional runs. After completion of all scans, participants filled out several personality questionnaires, were debriefed, and paid.

fMRI emotional appraisal task. Subjects were told that they would be presented with different scenarios. As illustrated in Figure 1, they were asked to read the scenario, and then to rate their emotional response (disgust-repulsion vs attractiveness-praiseworthiness). Participants were explicitly told to consider only their own emotional responses, and that although they might have thoughts about the overall social costs and benefits of the scenario, they should ignore them in making their rating. Participants had up to 10 s to respond, using a 4-point rating scale (1 = "Extremely appalling" to 4 = "Extremely appealing"). The right-to-left orientation of the rating scale was counterbalanced across subjects, but was kept consistent within subjects.

Table 1. Example stimuli[a]

Scenario	Mean emotional rating	Mean utilitarian rating
Stand-alone evil deeds		
Forcibly remove one kidney from an elderly person who is dying	−1.07	−0.57
Forcibly remove all organs from a young healthy child	−1.36	−1.29
Put out cigarette butts on a captured terrorist's face	−0.86	−0.43
Push large man in front of a runaway bus, killing him instantly	−1.29	−1.21
Push large man in front of a runaway bus, dismembering and killing him slowly via blood loss	−1.43	−1.25
Stand-alone greater goods		
Save life of another person by transplanting organs	1.04	0.79
Save life of U.S. president by transplanting organs	1.04	0.96
Gain information that saves Los Angeles from a terrorist nuclear weapon	1.36	1.46
Prevent injury of five pedestrians by a runaway bus	1.29	0.86
Prevent killing of five schoolchildren by runaway bus	1.25	1.11

[a]Ratings were made on a scale from 1 to 4. Means reported in the table were adjusted to have a range running from −1.5 to 1.5 to better distinguish negatively and positively valenced items.

Subjects rated 62 different scenarios (31 greater goods and 31 evil deeds; for examples, see Table 1). The scenarios were designed to be comprehensible both when presented separately as stand-alone acts during the appraisal tasks, and when presented in pairs during the moral judgment task. Trials were separated by a random intertrial interval (ITI) ranging from 2 to 7 s (mean = 4.3 s).

fMRI utilitarian appraisal task. This task was identical to the emotional appraisal task, except for the instructions given to subjects. Subjects were told to consider only the overall costs and benefits of the scenarios described (including the effects not just for those directly involved, but also for those affected indirectly, such as society as a whole). Participants were also told that, while they might feel strong emotions in response to the scenarios, they should consider only the costs and benefits in making their ratings. Participants responded using a similar 4-point rating scale (1 = "Extremely costly" to 4 = "Extremely beneficial"). The same 62 scenarios were used in both tasks.

fMRI overall moral judgment task. Subjects had to evaluate the overall moral appropriateness of moral tradeoffs, which were constructed from combinations of one stand-alone evil deed and one greater good previously rated by the subject (e.g., "Waterboard a captured terrorist" and "Save Los Angeles from terrorist nuclear attack"). As illustrated in Figure 2, the two scenarios were presented on opposite sides of the screen, with the side of the greater good and evil deed randomized across trials. Subjects were asked to provide a moral value judgment for the combined option by providing a rating of how appropriate it would be to perform the evil deed to obtain the greater good (1 = "Extremely inappropriate" to 4 = "Extremely appropriate"). They were told to assume that the only way to achieve the greater good was to perform the evil deed and that, if they judged it appropriate (rating = 3 or 4), both the evil deed and the greater good would occur exactly as described; but that if they judged it inappropriate (rating = 1 or 2), neither could occur. Participants had up to 15 s to respond in each trial; 132 moral tradeoffs were constructed from particular combinations of the 61 stand-alone acts, such that the

pair of one evil deed and one greater good always made a sensible complete scenario. The same set of 132 pairs was presented to all subjects, in an order randomized across subjects, and divided evenly across three scanning runs. The random ITI ranged from 3 to 5 s (mean = 4 s).

MRI data acquisition. Functional imaging data were collected using a Siemens 3.0 T Tim Trio MRI scanner to acquire gradient echo T2*-weighted EPI images at a transverse-to-coronal oblique tilt of −20 degrees using a 32-channel phased array coil. Each volume comprised 47 axial slices, acquired in an ascending manner using the following parameters: 30 ms TE; 192 mm FOV; 3 mm isotropic voxel resolution; and 2.5 s TR. Because participants were allowed to respond freely to each scenario up to a maximum time limit, the number of TRs differed in each functional run (range = 169–278 for the appraisal scans, 111- 205 for the moral judgment scans). We discarded the first two volumes of each functional run to allow for scanner equilibration. We also acquired a whole-brain high-resolution T1-weighted structural scan using 1 mm isotropic voxels for coregistration with the participant's mean EPI images.

MRI data preprocessing. We performed image analysis using SPM8 software (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London). Images were corrected for slice acquisition time within each volume, motion corrected with realignment to the last volume, spatially normalized to the standard MNI EPI template, and spatially smoothed using an isotropic Gaussian kernel with a FWHM of 8 mm. Intensity normalization and high-pass temporal filtering (filter width = 128 s) were also applied to the data.

MRI data analysis. We estimated several first-order autoregressive GLMs of BOLD response to address the various questions posed by the study. Each GLM was estimated in three steps. First, we estimated the model separately for each individual. Second, we calculated contrast statistics at the individual level. Third, we computed second-level statistics by performing one-sample *t* tests on the single-subject contrast coefficients.

GLM 1. This model was used to identify regions that parametrically encode the strength of emotional or utilitarian appraisals assigned to stimuli during the two appraisal tasks. The model had six regressors of interest. R1 is an indicator function for the decision period during emotional appraisals. R2 is an indicator function for the decision period during utilitarian appraisals. Both regressors last from trial onset to response and thus have a duration equal to the response time on that trial. R3 and R4 are parametric modulators of each indicator function with the explicit rating provided on each trial, that is, emotional ratings during the emotional appraisal task (R3), and utilitarian ratings during the utilitarian appraisal task (R4). R5 and R6 are parametric modulators of each indicator function giving the rating for the nonrequested appraisal, that is, the utilitarian ratings that the participant provided, but for the scenarios while shown during the emotional appraisal trial (R5), and vice versa (R6). Nonrequested ratings were orthogonalized to the explicitly requested ratings (i.e., R5 was orthogonalized to R3, and R6 was orthogonalized with respect to R4). Missed response trials were modeled as a separate regressor, with a duration of 10 s. All regressors were convolved with a canonical form of the hemodynamic response function. The model also included motion parameters and session constants as regressors of no interest.

Several subject-level contrasts were calculated and submitted to a group random-effects analysis: C1, the contrast R1-R2 characterized regions with stronger average responses during emotional versus utilitarian appraisal tasks; C2, we looked for areas where BOLD response correlates with emotional appraisals using the parameter estimate for R3; C3, we looked for areas where BOLD response correlates with utilitarian appraisals using the parameter estimate for R4; C4, we computed the difference R3-R4 to look for regions that reflected either utilitarian or emotional appraisals, but not both; and C5, because all responses for C2 were positive, but all responses for C3 were negative, we used the contrast R2-R3 to look for regions with a different absolute size in the responses for utilitarian and emotional appraisals.

A variant of this model that included and controlled for reaction time (RT) as a parametric modulator yielded nearly identical results, and so is not discussed further here.

GLM 2. This model was used to test whether the regions identified in GLM 1 also represent emotional and utilitarian appraisals during the moral judgment task, even though such representations are not explicitly required. This model was estimated using only the moral judgment trials. It had the following regressors of interest: R1 consisted of an indicator function beginning at trial onset and ending when the subject made a response. R2 was a parametric modulator of R1 with the sum of the emotional ratings for the evil deed and greater good shown on that trial, based on the participant's responses during the appraisal tasks. R3 was a second parametric modulator of R1 with the sum of the utilitarian ratings, computed in the same manner. The second parametric modulator was orthogonalized to the first one. Here, and below, all omitted details are as in GLM 1.

We analyzed the results of this model using an ROI approach designed to test whether the same areas from GLM 1 also represented emotional and utilitarian concerns during the moral tradeoff task. This was done by computing the average estimated coefficient for R2 and R3, separately for each subject within the ROIs functionally defined using GLM 1 (see Fig. 3), and then comparing them using *t* tests.

For robustness, we also ran a similar model with the order of parametric regressors reversed, causing a change in the order of the orthogonalization. This model yielded largely similar results, so is not discussed further.

GLM 2B. This model was used to test whether encoding of utilitarian and emotional appraisals during the moral judgment task was choice-dependent (e.g., stronger if a participant ultimately judged a tradeoff appropriate instead of inappropriate). The model was identical to GLM 2, with the exception that all regressors of interest were computed separately based on whether the participant ultimately decided that a tradeoff was appropriate (i.e., response = 3, 4 during the tradeoff task) or inappropriate (i.e., response = 1, 2).

GLM 3. This model was used to identify regions that encode the overall moral value ratings made during the moral judgment task. It had two regressors of interest: R1 was an indicator function for moral judgment trials, with a duration from trial onset to trial response; and R2 was a parametric modulator of R1 with the appropriateness rating for that trial. Missed response trials were modeled as a regressor of no interest (duration = 15 s). To identify regions associated with the computation of overall moral value, we computed a one-sample *t* test against zero using the single-subject estimated coefficients from regressor R2.

GLM 4. This model was used to identify regions that encode the overall moral value ratings only at specific times within a decision trial (e.g., only before response). The model had the following regressors of interest: R1 was an indicator function for the first 2 s of each moral judgment trial; R2 was an indicator function for the last 2 s before response of each moral judgment trial; R3 was a parametric modulator of R1 with the appropriateness rating for the trial; and R4 was a parametric modulator of R2 with the appropriateness rating for the trial. Parametric modulators for the early and late periods for a given trial are identical, but the correlation between these regressors when convolved with the hemodynamic response is quite low because of the multisecond lag between the beginning and end of each trial. Missed response trials were modeled as 2 s boxcar functions representing the first 2 s after trial onset, and another representing the final 2 s of the trial (13–15 s after trial onset). All other details are as in GLM 1. These regressors were used to estimate the following contrasts: C1, correlations with overall appropriateness early in the trial (R3); C2, correlations with overall appropriateness just before the response (R4); and C3, the interaction of moral judgment with time (R4-R3).

We report regions as significant if they passed whole-brain cluster correction (WBC) at $p < 0.05$ as implemented in SPM8 (Worsley et al., 1996), using a per-voxel threshold of $p < 0.001$. Regions are also reported if they survived cluster-level small volume correction (SVC) at $p < 0.05$ within the following three a priori anatomically defined regions of interest:

1. A vmPFC ROI that included all voxels within the bilateral anterior cingulate cortex, rectus, and medial orbitofrontal gyrus from the AAL atlas and inferior to $z = 0$ (2733 voxels). This region encompasses the peak voxels related to value computation in several independent studies, including a region of anterior vmPFC identified by a meta-analysis of moral decision-making (Bzdok et al., 2012).

2. A temporoparietal junction (TPJ) ROI that included bilateral angular and superior temporal gyrus, posterior to $y = -40$ (1975 voxels). This region encompasses peaks of activation from several studies of social cognition and Theory of Mind (Gallagher and Frith, 2003; Decety and Jackson, 2006; Saxe and Powell, 2006).
3. An amygdala ROI consisting of right and left amygdala as defined in the AAL atlas (128 voxels), which has been associated with emotional ratings in previous studies of moral judgment (Shenhav and Greene, 2014).

These ROIs were defined anatomically using the WFU PickAtlas plugin for SPM (http://fMRI.wfubmc.edu/software/PickAtlas), with a dilation of 3 mm to ensure full coverage of the area.

Finite impulse response (FIR) analyses. To examine the time course of neural activation, we estimated several FIR models of the BOLD response during the moral judgment task. These analyses were designed to shed light on the timing by which the emotional appraisals, utilitarian appraisals and overall moral value arise in vmPFC.

FIR 1. This was designed to examine the dynamics of vmPFC representations of overall moral value. To do this, we defined an ROI in the vmPFC based on the set of voxels correlating with overall appropriateness in the final 2 s before the response, thresholded at $p < 0.001$ uncorrected (see Fig. 5). From this ROI, we extracted the average raw BOLD time signal at each time point, removing both the mean and variance associated with motion regressors using standard SPM functions. The resulting time course was then up-sampled using spline interpolation into 10 time bins per TR (250 ms per bin), in a manner similar to the approach used in several other studies examining the neural time course of information representation (Boorman et al., 2009; Hutcherson et al., 2012; Chau et al., 2014). We then estimated a FIR linear regression model with two regressors for each time point: a constant and the appropriateness rating provided in that trial. The estimated regression coefficients for the appropriateness ratings at each point in time were used to construct the time course plots and tests in Figure 5.

FIR 2. This model was similar to the previous one, with the exception of the location of the time bins. These were placed at 250 ms increments beginning 13 s before and continuing for 7 s after response. This allowed a complete visualization of hemodynamic responses aligned to the responses (in contrast, in FIR 1 the time course is aligned to the trial onset).

FIR 3. This model was used to examine the dynamics of the vmPFC responses to emotional and utilitarian appraisals. The model was similar to FIR 1, with the exception that it included two parametric regressors for each time bin: one for the total emotional appraisal of the combined stimulus and one for the total utilitarian appraisal (for details, see GLM 2).

FIR 4. This model was used to investigate whether parametric appraisal representations were more clearly locked to stimulus onset (as might be predicted by an attribute integration account), or to the responses themselves. It was identical to FIR 3, with the exception that the time window of interest ran from 13 s before response until 7 s afterward.

In all of these models, we used permutation tests to assess statistical significance for each of the above FIR analyses while correcting for the number of comparisons conducted for the time-points of interest. Specifically, for each model, we reran the analysis 1000 times, using a randomly permuted order for the trial-level data (i.e., appropriateness ratings or attribute ratings). The results at each time-point were thresholded at $p < 0.05$ uncorrected, and corrected significance was determined using the number of consecutively significant time-points required to protect against false positives at the $p < 0.05$ level during the 6 s window approximating the stimulus evaluation period, adjusting for the hemodynamic lag (i.e., time-points 17–41).

Temporal variation in functional connectivity. We hypothesized that each distinct appraisal is passed to vmPFC to be integrated into an overall moral value. This predicts that there should be an increase in functional connectivity between the systems computing the appraisals and the value-related vmPFC during the moral judgment trials. This prediction is usually tested in fMRI using a psychophysiological interaction analysis. However, standard psychophysiological interaction analyses are problematic in our data because the results of the FIR analyses suggested the possibility that emotional and utilitarian appraisals are represented in vmPFC at different times. This in turn implies that the connectivity between the vmPFC and lower-level regions might show considerable variability over time within the trial. We addressed this problem by using a variant of the FIR analyses that allowed us to examine the temporal profile of connectivity between the vmPFC and regions computing the separate appraisals.

To do this, we extracted BOLD time courses from three ROIs correlating with emotional appraisals in GLM 1 (i.e., left superior temporal gyrus [STG], right insula, and anterior cingulate cortex [ACC]), and three regions correlating with utilitarian appraisals (i.e., right TPJ, left ventrolateral prefrontal cortex [vlPFC], and the dorsomedial prefrontal cortex [dmPFC]), using the procedure described in the FIR analyses above. We then conducted six different FIR linear regression models for vmPFC responses, one for each source ROI associated with the computation of emotional or utilitarian appraisals. These models were identical to FIR 3, except that they also included two additional variables: the up-sampled activity of the source ROI (e.g., ACC or dmPFC) during the time points of interest, and the up-sampled activity averaged over the whole brain to control for nonspecific BOLD effects. The estimated regression coefficients for the source ROI regressor, which measure the correlation between vmPFC and the source ROI at each time point, were subjected to a one-sample t test against 0 at the group level. Permutation tests similar to those described for the FIR analyses determined corrected levels of significance.

Within- and between-network regional interaction analyses. We performed the following two analyses to investigate the extent to which the areas involved in computing the different appraisals operate independently, or demonstrate competitive inhibition.

First, we computed FIR-based connectivity analyses between the six ROIs that correlate with either emotional or utilitarian appraisals. Each analysis was similar to the ones described in the previous section, except that now it was performed between each pair of the appraisal ROIs, instead of the vmPFC. As a summary measure of connectivity between each pair of regions, we averaged all connectivity coefficients from the average deliberation period duration of the moral judgment task, adjusting for the hemodynamic lag (i.e., the period from 4 to 10 s after onset of the stimulus). Results are displayed in Figure 7*A*.

Second, we asked whether the appraisal representations in each of the six ROIs (i.e., strength of encoding emotional ratings for emotion regions and utilitarian ratings for utilitarian regions) were correlated either between areas within a network or between areas across networks. For emotional appraisal ROIs, we computed the strength of emotional appraisal coding during the moral judgment task (R2 from GLM 2), averaging over all voxels within each region separately. For utilitarian appraisal ROIs, we computed the strength of utilitarian appraisal coding (R3 from GLM 2). We then computed the Pearson correlation coefficient between these measures for each pair of regions. Results are displayed in Figure 7*B*.

Results

We begin with an overview of the logic of the study. Subjects performed three tasks while we measured BOLD responses with fMRI. First, subjects were shown one morally relevant action at a time and were asked to provide an emotional appraisal for it using a 4-point scale (1 = "Extremely Appalling" to 4 = "Extremely Appealing"), considered independently from the action's overall social utility. Second, subjects were shown the same stimuli and had to provide a utilitarian appraisal describing the social benefit of the proposed moral act, considered independently from emotional response (1 = "Extremely Costly" to 4 = "Extremely Beneficial"). The proposed moral acts used in these two tasks ranged from problematic actions, such as "forcibly removing organs from young children" to more desirable actions, such as "saving the life of the U.S. president" (Fig. 1*A*; for sample stimuli, see Table 1). For convenience, we refer to these as "evil deeds" and "greater goods," respectively. The order of the two appraisal tasks was counterbalanced across subjects. Third, fol-

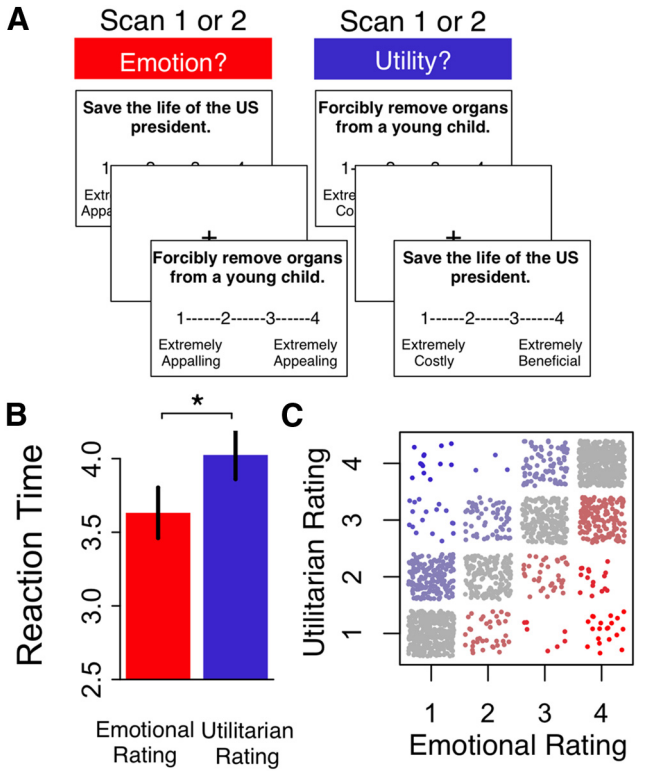

Figure 1. Emotional and utilitarian appraisal tasks for stand-alone acts. ***A***, Trial structure. ***B***, Average response times (± SE) to make emotional (red) or utilitarian (blue) appraisals. *$p <$ 0.05. ***C***, Relationship between emotional and utilitarian ratings for the same act. Each point indicates the ratings made by a single subject for a single act, with red points indicating a higher emotional rating for the same act, blue points indicating a higher utilitarian rating, and gray points indicating the same rating.

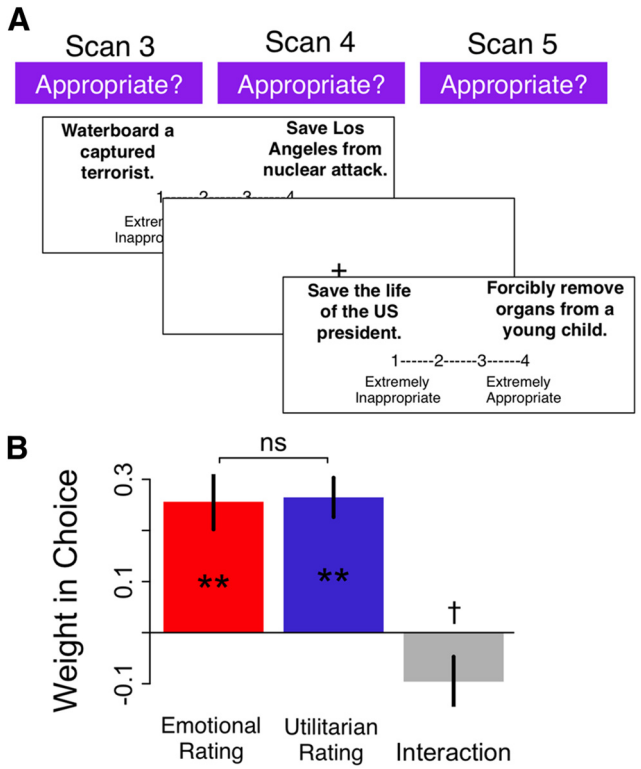

Figure 2. Overall moral judgment task for combined acts. ***A***, Trial structure. ***B***, Average influence on overall moral values of overall emotional appraisals (red), overall utilitarian appraisals (blue), or their interaction (gray). Error bars indicate SEM. **$p < 0.001$. †$p = 0.06$. ns, Not significant.

lowing the appraisal tasks, subjects completed an overall moral judgment task (Fig. 2*A*). Subjects were shown moral dilemmas that were constructed by combining one of the greater goods and one of the evil deeds that they had rated previously, and were asked to provide an overall moral value based on the appropriateness of performing the evil deed to achieve the greater good (1 = "Extremely Inappropriate" to 4 = "Extremely Appropriate").

We had several hypotheses about how subjects made overall moral judgments, motivated by the idea that the underlying architecture might resemble the one that has been shown to be at work in some simple economic choice paradigms (Fehr and Rangel, 2011; Lim et al., 2013; Rangel and Clithero, 2014). First, we hypothesized that the emotional and utilitarian appraisals would be computed independently in separate sets of regions, during both the explicit appraisal tasks and the overall judgment task. Second, we hypothesized that the overall judgment value signal is encoded in the same vmPFC regions that have been shown to compute value in a multitude of decision tasks (Bartra et al., 2013; Clithero and Rangel, 2014). Third, we hypothesized that the overall value signal is computed by integrating attribute information, such as the emotional appraisals and the utilitarian appraisals, within the vmPFC.

Several aspects of the design are worth emphasizing. First, the appraisal tasks allow us to obtain type-specific appraisals for each component act, to independently identify regions that encode each type of appraisal, and to test whether these regions specialize in representing one type of appraisal. Second, we can estimate the emotional and utilitarian appraisals of the combined acts, by adding together the appraisals of the greater good and evil deed elicited during the appraisal tasks. This allows us to look for regions encoding emotional and utilitarian appraisals during the overall moral judgment task, even though such ratings were not explicitly provided by the subject. Third, by identifying systems associated with different appraisals, we can test the hypothesis that emotional and utilitarian appraisals are computed mostly independently by looking for interactions between their neural representations. For example, some models propose that the emotional and utilitarian systems compete through mutual inhibition, with the winner taking control of behavior (Greene et al., 2004; Cushman, 2013), which predicts a negative correlation between the two representations.

Behavior: appraisal tasks

Participants were faster to provide emotional appraisals than utilitarian appraisals ($M_{Emotional} = 3632$ ms, SE = 164; $M_{Utilitarian} = 4025$ ms, SE = 171; paired $t_{(27)} = 2.5$, $p = 0.02$; Fig. 1*B*), consistent with the hypothesis that they might be supported by different processes. Despite this RT difference, emotional and utilitarian ratings were strongly correlated within subjects (Fig. 1*C*; mean $r = 0.78$). However, this high correlation was in part an artifact of collapsing over evil deeds and greater goods. Computing the correlation between emotional and utilitarian ratings only within evil deeds, or only within greater goods, yielded more modest relationships (mean r evil deeds = 0.42, mean r greater goods = 0.28).

Behavior: moral judgment task

Participants judged the combination of evil acts and greater goods as appropriate (i.e., a value of 3 or 4) on 56 ± 4% (mean ± SD) of trials. For every subject, we estimated a linear regression of the overall moral value on the overall emotional appraisal (given

by the sum of the subject's emotional appraisals for the evil deed and greater good on that trial), the overall utilitarian appraisal (defined similarly), and their interaction. This regression provides a behavioral test of the hypothesis that the overall moral value reflects the integration of emotional and utilitarian appraisals, among other types of information. The within-subject correlation between the two regressors was low to moderate (mean $r = 0.33$), which means that the regression is able to estimate the contribution of both types of appraisals. Estimated subject-level coefficients were submitted to one-sample *t* tests to determine the significance of each factor. As shown in Figure 2*B*, both emotional and utilitarian appraisals contributed significantly to overall judgments ($M_{Emotional} = 0.256 \pm 0.054$, one-sample $t_{(27)} = 4.66$, $p < 0.001$; $M_{Utilitarian} = 0.265 \pm 0.038$, one-sample $t_{(27)} = 7.22$, $p < 0.001$), with no significant difference between them (paired $t_{(27)} = 0.12$, $p = 0.91$). We observed a marginally significant interaction effect ($M_{Interaction} = -0.096 \pm 0.049$, one-sample $t_{(27)} = 1.97$, $p = 0.06$, two-tailed), which had a very minor contribution to the variance explained by the regression (mean increase in R^2 of 0.02). Together, these results are consistent with the hypothesis that overall moral value reflects the integration of emotional and utilitarian appraisals, with a weak or nonexistent interaction between them.

Average neural responses in the appraisal tasks

The key neural hypotheses of the study involve the existence of areas with neural responses that parametrically encode the emotional appraisals, the utilitarian appraisals, or the overall moral values. However, to facilitate comparison with previous literature in moral judgment, which has often focused on cross-condition comparison, in this section we compare the average BOLD responses during the emotional and utilitarian appraisal tasks (for details, see GLM 1). We found that supplementary motor area ($p < 0.05$, WBC, peak at 9, 26, 31) and left amygdala ($p = 0.04$, SVC, peak at $-24, 2, -23$) exhibited greater average BOLD responses during the emotional value rating task. No regions had significantly greater average activation during the utilitarian appraisal task.

Appraisal signals during the appraisal tasks

We began the primary analyses by looking for regions in which the BOLD responses were parametrically correlated with the emotional or utilitarian appraisals during the appraisal tasks (for details, see GLM 1). In these tasks, subjects are expected to compute each respective type of appraisal because they have to report it. This allows us to independently identify areas associated with representing each type of appraisal, without assuming or requiring that such appraisals be used during the moral judgment task (e.g., some subjects might be able to report utilitarian appraisals when asked, but might not compute or use this information to make moral judgments).

We found a dissociation between areas that reflected the emotional and utilitarian appraisals (Fig. 3; Table 2). Emotional appraisals correlated positively with BOLD responses in anterior cingulate cortex (ACC; $p < 0.05$, whole-brain corrected), right superior temporal gyrus (STG; $p < 0.05$, WBC), and, marginally, in right mid-insula ($p < 0.07$, WBC) and left STG ($p < 0.08$, WBC). No such correlation was observed in the amygdala, even at much lower significance levels, nor did we observe a significant correlation between amygdala and emotional appraisals when analyzing negative or positive emotional scenarios separately (all *p* values > 0.05 uncorrected). In contrast, utilitarian appraisals correlated with BOLD responses in the dorsomedial prefrontal cortex (dmPFC; $p < 0.05$, WBC), ventrolateral prefrontal cortex (vlPFC; $p < 0.05$, WBC), and right temporoparietal junction (TPJ; $p < 0.05$, SVC). Intriguingly, and in contrast to emotional appraisals, representations in each of these regions correlated negatively with utilitarian appraisals (i.e., responding more strongly for costly actions than beneficial ones). No correlation with utilitarian appraisals was observed in areas of the dorsolateral prefrontal cortex that respond to selection of utilitarian options during difficult personal moral dilemmas (Greene et al., 2001; Greene et al., 2004), even at a liberal threshold of $p < 0.05$ uncorrected.

These findings suggest that emotional and utilitarian appraisals are computed in dissociable regions. To better understand the specificity of response to emotional and utilitarian appraisals, we performed two additional whole-brain analyses (Table 2). First, we looked for areas with a significant difference between the coefficients for utilitarian and emotional appraisals (i.e., where the difference $\beta_{Util} - \beta_{Emot}$ was significantly different from 0). This analysis identified only the right TPJ ($p < 0.07$, WBC; $p < 0.05$, SVC, peak $x, y, z = 45, -61, 31$), where the difference was driven by a negative correlation with utilitarian appraisals that was significantly different from the association with emotional appraisals. Unfortunately, interpretation of this contrast is complicated by the fact that emotional appraisals were generally represented positively, whereas utilitarian appraisals were generally represented negatively. Given this, we performed a second more demanding analysis, in which we looked for regions where the negative response to the utilitarian appraisals was significantly different from the positive response to the emotional appraisals (i.e., where the difference $\beta_{Util} - \beta_{Emot}$ was significantly different from 0). This analysis identified regions of both dmPFC and vlPFC where the negative response to the utilitarian appraisal was stronger than the positive response to the emotional appraisal, as well as a region encompassing both ACC and the ventral striatum where the positive response to the emotional appraisal was stronger than the negative response to utilitarian appraisals (all *p* values <0.05, WBC). For illustrative purposes only, given the nonindependent nature of the analysis, Figure 3 depicts the average estimated regression coefficients for the two types of ratings in each of the ROIs identified by these results.

The results in this section suggest that emotional appraisals of moral acts are dominant in ACC and that the opposite is true in dmPFC and vlPFC. Figure 3 suggests that right mid-insula and the left STG may also exhibit dominant representations of emotional appraisals and that the right TPJ may exhibit a dominant representation of utilitarian appraisals, but the dissociation tests for these regions were inconclusive within the statistical power offered by this dataset.

Appraisal signals during the overall moral judgment task

The next step in the analysis asks whether the areas encoding emotional and utilitarian appraisals during the appraisal tasks, when subjects are required to compute them, also encode them in the same way during the moral judgment task, when they are not explicitly asked to do so. This provides a test of the hypothesis that emotional and utilitarian appraisals are computed during the process of making a moral judgment. We used the areas identified in the previous analysis to functionally define ROIs associated with appraisal-specific computations, and then asked whether these regions still reflected these representations during the overall moral judgment of tradeoffs (for details, see GLM 2). For the most part, we found that the appraisal-specific representations persisted in the moral judgment task, and with the same sign (Fig. 4). Right insula, left STG, and ACC responses all corre-

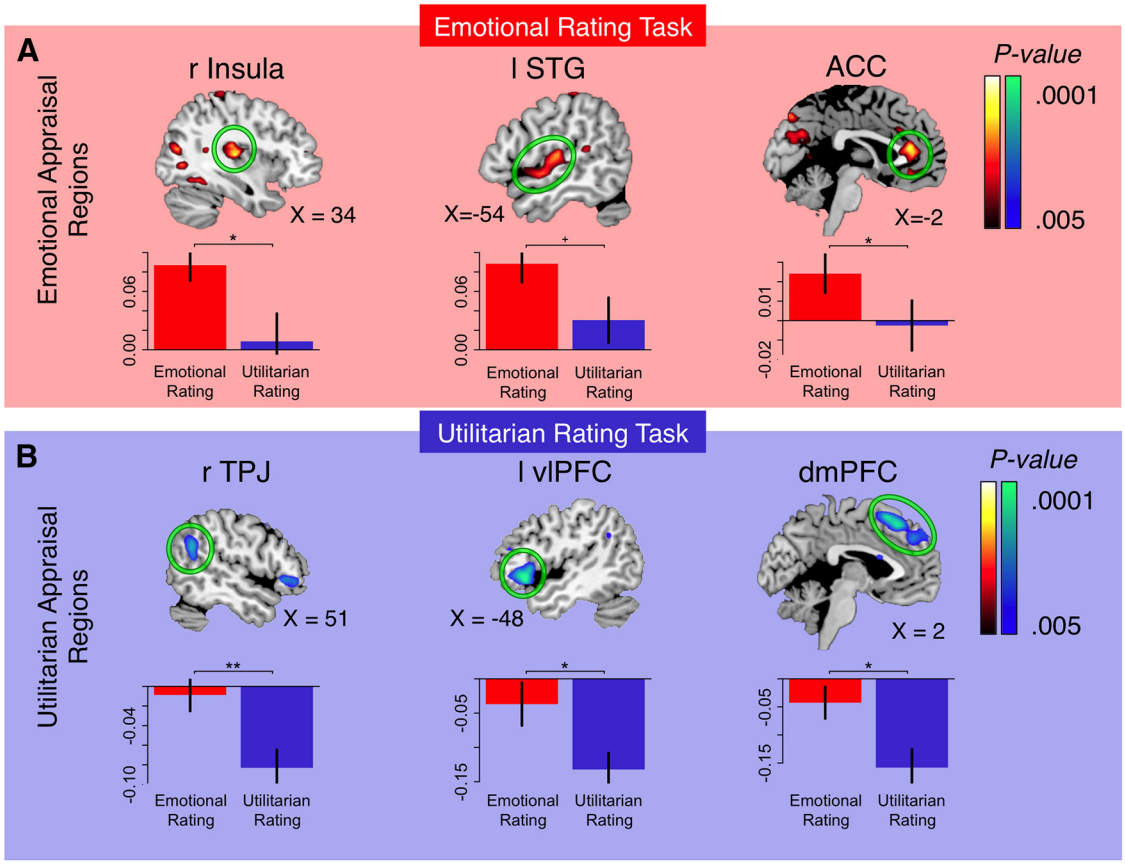

Figure 3. Neural correlates of emotional and utilitarian appraisals for stand-alone acts. Different regions correlated with the two types of appraisals during the emotional appraisal task (***A***) and the utilitarian appraisal task (***B***). Hot colors represent a positive correlation. Cold colors represent a negative correlation. Images thresholded at $p < 0.001$, uncorrected. Bar plots represent correlation with emotional and utilitarian appraisals and are shown for visualization purposes only. Error bars indicate SEM. *$p < 0.05$. **$p < 0.01$.

Table 2. Neural correlates of emotional appraisals, utilitarian appraisals, and overall moral judgments[a]

Region	BA	Volume	Z	x	y	z
Correlation with emotional appraisals						
Right Anterior cingulate cortex	24/32	124	4.42	6	35	16
Right Superior temporal gyrus	22	70	3.85	48	−10	−5
Right Parahippocampal gyrus	36	54	4.04	24	−34	−14
Right Mid-insula	13	48	3.77*	36	−19	10
Left Occipital cortex	18	47	4.04*	−21	−85	13
Left Superior temporal gyrus	22	46	3.89*	−54	−13	−2
Correlation with utilitarian appraisals						
Left Dorsomedial prefrontal cortex	8	200	4.27	−6	17	58
Left Ventrolateral prefrontal cortex	45/47	110	4.46	−48	26	−5
Right Temporoparietal junction	40	30	3.76**	48	−55	25
Emotional appraisals (positive) > utilitarian appraisals (negative)						
Left Ventral striatum		273†	6.04	−9	22	−9
Right Anterior cingulate cortex	24		3.92	3	35	7
Utilitarian appraisals (negative) > emotional appraisals (positive)						
Right Dorsomedial prefrontal cortex	8	112	4.76	9	35	52
Left Ventrolateral prefrontal cortex	45/47	90	4.11	−39	26	−8
Correlation with overall moral value (full period)						
No regions significant						
Correlation with overall moral value (first 2 s)						
No regions significant						
Correlation with overall moral value (last 2 s)						
Right Ventromedial prefrontal cortex	10	17	4.11††	0	53	−2

[a]Regions are reported at $p < 0.05$, whole-brain corrected, unless otherwise noted.

*$p < 0.08$, whole-brain corrected, reported for completeness. **$p < 0.05$, small-volume corrected within a bilateral anatomical mask of the temporoparietal area. †Part of larger cluster, reported for completeness. ††$p < 0.05$, small-volume corrected within a bilateral anatomical mask of vmPFC.

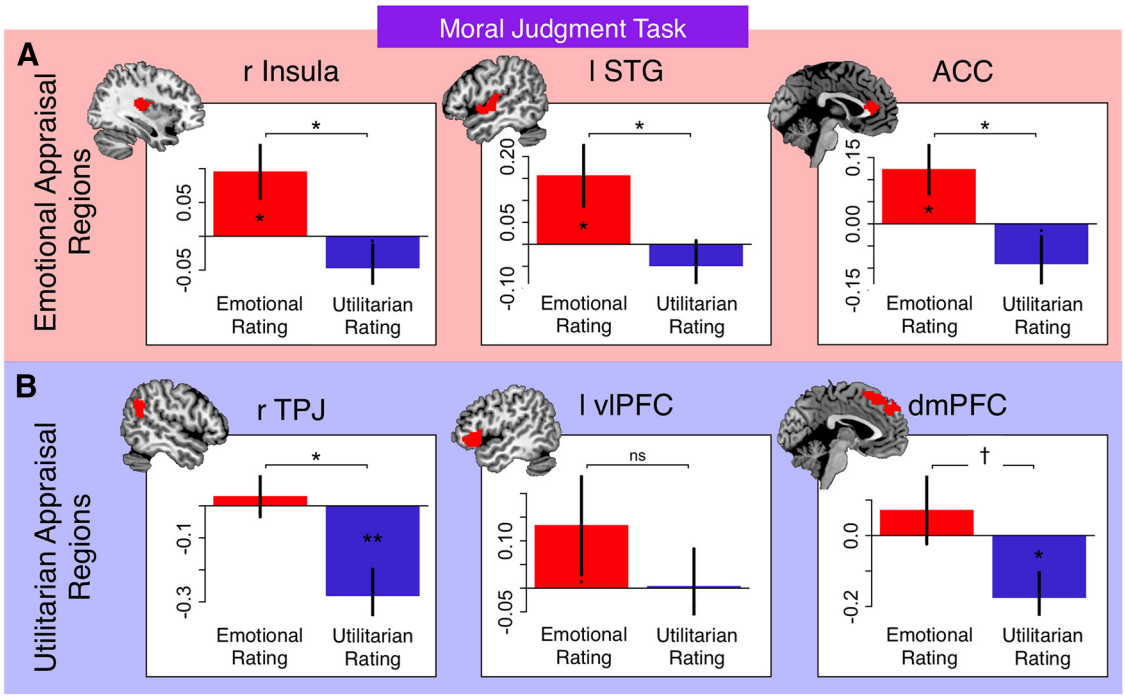

Figure 4. Representations of emotional and utilitarian appraisals persist in the moral judgment task. ***A***, Response in regions associated with emotional appraisals during the appraisal tasks correlated with the overall emotional appraisal during the moral judgment task but did not correlate with the overall utilitarian appraisal. ***B***, Response in regions associated with utilitarian appraisal during the appraisal tasks correlated with overall utilitarian appraisal of the tradeoff during the judgment task but did not correlate with overall emotional appraisals. Regional masks used to extract these values are shown in the upper corner of each graph and were defined using the set of voxels that correlated at a significance level of $p < 0.001$ with emotional or utilitarian appraisals in GLM 1. Error bars indicate SEM. *$p < 0.05$. **$p < 0.01$. †$p = 0.07$. ns, Not significant.

lated significantly and positively with the overall emotional appraisal (i.e., the sum of the emotional ratings for the evil deed and greater good on each trial, all p values <0.05) but did not correlate with the overall utilitarian appraisal (i.e., the sum of the utilitarian ratings for the evil deed and greater good on each trial). In contrast, right TPJ and dmPFC (though not the vlPFC) correlated significantly and negatively with the overall utilitarian appraisal (all p values <0.05) but did not correlate with overall emotional value. In addition, we found no significant differences in the strength of emotional or utilitarian attribute coding in any of these regions as a function of whether the participant ultimately judged the tradeoff as appropriate or inappropriate (for detils, see GLM 2B).

These results suggest that emotional and utilitarian appraisals are represented in largely specialized regions during moral judgment and that these are the same regions that encode explicitly requested appraisals. In addition, the fact that the appraisal representations are not affected by the moral judgment made (i.e., appropriate vs inappropriate) is consistent with the hypothesis that the appraisals are computed independently. It is less consistent with a model in which the two representations mutually inhibit each other in a winner-take-all competition.

Overall moral values are represented in vmPFC

Next, we tested the hypothesis that overall moral values are represented in vmPFC during the moral judgment task, as might be expected from studies of simple value-based choice. We did this by looking for areas in which BOLD responses correlated with overall moral value, either averaged over the whole decision period or in the moments just before the participant made a response (for details, see GLM 3 and GLM 4). Although no regions significantly correlated with moral value when analyzing the decision-period as a whole, we observed a single area exhibiting a significant positive correlation with the overall moral value in the 2 s before response, located in the anterior vmPFC ($p < 0.05$, SVC; peak at 0, 53, −2; Fig. 5*A*). In contrast, there were no regions that correlated significantly with overall moral value in the first 2 s after trial onset. Time course analyses using an FIR model of responses within this vmPFC region (Fig. 5*B*) confirmed that the overall moral value signal emerged shortly before a participant made the final response. Furthermore, when responses were analyzed time-locked to the onset of the trial, this region showed a significant but prolonged and smeared out correlation with the overall moral value, centered just before and during the average response time. In contrast, when the analysis was time locked to the response itself, vmPFC activity exhibited a sharp increase in correlation with the overall moral value just before a response was made. Together, these results are consistent with the hypothesis that an overall moral value signal arises in vmPFC just before the time of response.

Attribute value representations in the integrative vmPFC region

Next, we performed two separate analyses to test the hypothesis that the overall moral value signal in vmPFC reflects the integration of emotional and utilitarian appraisal information, and to probe how this integration evolves over the course of the decision.

We first asked whether the vmPFC represented both emotional and utilitarian appraisals during the moral judgment task, as predicted by the hypothesis that the overall moral value encoded here reflects the integration of both types of appraisals. In particular, we analyzed the time course of responses in the region of vmPFC identified above, using an FIR model to determine

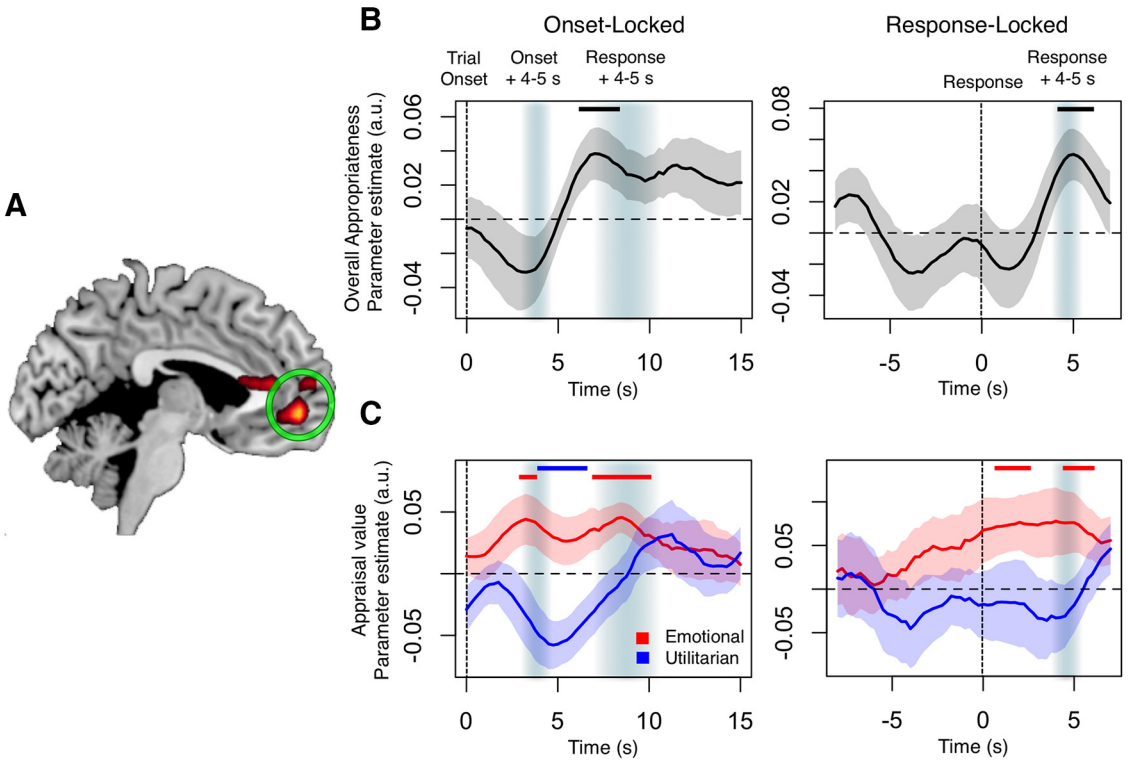

Figure 5. Neural correlates of overall value in the moral judgment task. ***A***, A region of the vmPFC correlated with overall moral value in the 2 s before response. ***B***, Time course of the overall moral value signal in this region, locked either to onset of the tradeoff (left), or the response (right). These time courses are included for display purposes only and do not represent independent tests. ***C***, Time course in vmPFC of the overall emotional appraisal (red), and overall utilitarian appraisal (blue), locked to onset (left) or response (right). Lighter shaded areas represent SEM. Lines at the top of the charts indicate time-points where the association was significant at $p < 0.05$, uncorrected for multiple comparisons. Gray vertical shading represents a visual aid for the approximate time of stimulus onset and response time, adjusted for the hemodynamic lag.

whether and when the activity in this ROI correlated with the overall emotional or utilitarian appraisals of each proposed moral action. Consistent with the integration hypothesis, we found a significant correlation with both emotional and utilitarian attributes (Fig. 5*C*). The correlation began with a rapidly emerging but brief sensitivity to the overall emotional appraisals (beginning at $t = 3.25$ s following stimulus onset and lasting 1.25 s at $p < 0.05$ uncorrected; $p = 0.1$ permutation corrected), a slightly delayed but longer period of negative correlation with overall utilitarian appraisals (beginning at $t = 4.25$ s and lasting 2.75 s at $p < 0.05$ uncorrected; $p = 0.02$ permutation corrected), and finally a return to a more sustained positive correlation with emotional appraisals (beginning at $t = 7.25$ s and lasting for 3 s at $p < 0.05$ uncorrected; $p = 0.006$ permutation corrected).

Figure 5 depicts the evolution of the two appraisal signals. It suggests that the earliest correlation with emotional and utilitarian appraisals in the vmPFC emerged ~2–3 s earlier in time than the correlation with overall moral value, consistent with the idea that the appraisals provide inputs to this area that are subsequently integrated into an overall moral value. Further bolstering this interpretation, the sensitivity of the vmPFC to emotional or utilitarian appraisals was more sharply tuned to the onset of the trial and was weaker and more smeared out when we repeated the analysis time-locked to the moment of response (Fig. 5*C*, right).

We next tested a corollary of the value integration hypothesis: if the information about emotional and utilitarian appraisals within vmPFC reflects inputs from appraisal-specific areas (e.g., insula, ACC, TPJ, or dmPFC), then the vmPFC should exhibit increases in functional connectivity with these source regions at the time of making a moral judgment. As described in Materials and Methods, this analysis is complicated by the fact that inputs from these specialized regions may arise in vmPFC at different times (as suggested by the RT and neural differences reported above). As a result, connectivity between the vmPFC and areas representing these attributes may vary considerably over the course of a trial. In this case, a standard connectivity analysis using psychophysiological interaction models is problematic because it assumes that the precise timing of a psychological period of interest is known and that connectivity during this period is constant. To sidestep this problem, we performed a different type of functional connectivity analysis. In particular, we extracted the time course of BOLD responses from each of the ROIs associated with either emotional or utilitarian appraisals and then performed an FIR-like analysis to estimate the correlation between activity in each these regions and the vmPFC at different points over the course of the moral judgment trials (for details, see Materials and Methods).

We found only two regions exhibiting significant connectivity profiles with vmPFC during overall moral judgment. For the ACC region associated with emotional appraisals (Fig. 6*A*), we observed a significant degree of correlation with the vmPFC during the whole trial (beginning at $t = 0$ and lasting 15 s at $p < 0.05$ uncorrected, $p < 0.001$ permutation corrected), with a temporal profile consisting of an earlier and a later peak that matched the temporal profile of vmPFC sensitivity to emotional value. For the dmPFC region associated with utilitarian appraisals (Fig. 6*B*), we observed a more temporally circumscribed period of connectivity that lined up with the moments during which the vmPFC displayed the most sensitivity to the utilitarian appraisal of the tradeoff (beginning at $t = 5$ s and lasting 2.5 s at $p < 0.05$ uncor-

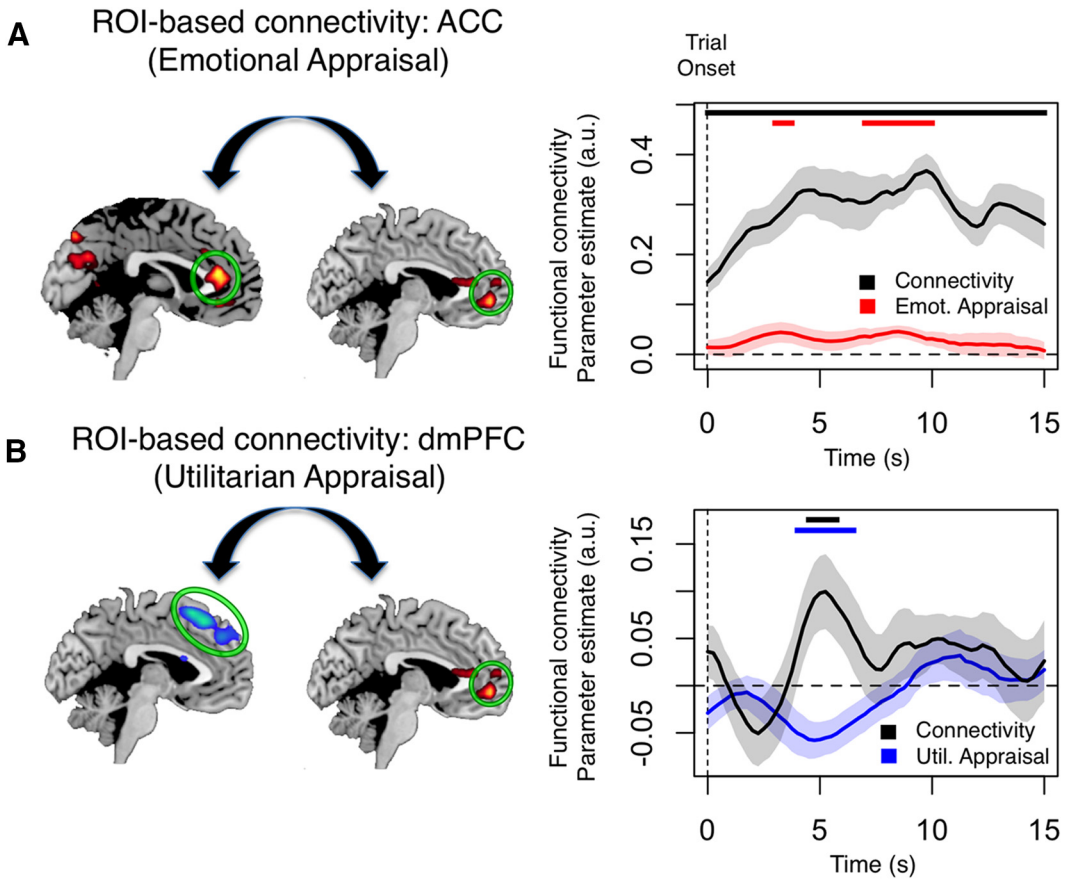

Figure 6. Time course of connectivity between vmPFC and appraisal-specific regions. ***A***, vmPFC and ACC connectivity (black line) was significant throughout the trial, with peaks around the times at which the vmPFC showed significant correlation to the overall emotional appraisals (red line). ***B***, vmPFC and dmPFC connectivity (black line) was significant during the same time-period that vmPFC became showed significant correlation to overall utilitarian appraisals (blue line). Lighter shaded areas represent SEM. Lines at the top of the charts indicate time-points where the association was significant at $p < 0.05$, uncorrected. Gray vertical shading represents a visual aid for the approximate time of stimulus onset and response time, adjusted for the hemodynamic lag.

rected; $p = 0.07$ permutation corrected). This pattern was unique to these two regions: the other ROIs exhibiting appraisal-specific responses, such as the TPJ or insula, showed either no significant functional connectivity with the vmPFC, or comparatively brief periods of connectivity that were not aligned with the timing of representations within the vmPFC (data not shown).

Together, the results in this section are consistent with the hypothesis that the overall moral value signal in vmPFC reflects the integration of emotional appraisals encoded in ACC, and of utilitarian appraisals encoded in dmPFC. Interestingly, the analyses are also suggestive that appraisal information might precede the appearance of the integrated overall moral value, although this result is more tentative.

Interaction between appraisal systems during the moral judgment task

The results described so far are consistent with a model of moral judgment in which separate systems independently compute emotional and utilitarian appraisals, which are then integrated in vmPFC to compute an overall moral value. In contrast with this hypothesis, several influential models have proposed that moral judgment involves inhibitory competition of the emotion and utilitarian appraisal regions, with the overall judgment being determined by the winner (Greene et al., 2001, 2004, 2008; Cushman et al., 2010). Our final analyses perform additional tests of the inhibitory model.

First, we looked for functional connectivity patterns within and across the six ROIs that correlate with emotional or utilitarian appraisals. To do this, we estimated the average pairwise connectivity between all these regions while participants made judgments in the moral tradeoff task (for details, see Materials and Methods). As shown in Figure 7*A*, we observed significant within-system connectivity (insula-ACC, $p = 0.03$ Bonferroni corrected for 15 separate comparisons; insula-STG, $p < 0.001$ corrected; TPJ-dmPFC, $p = 0.01$ corrected; IFG-dmPFC, $p = 0.04$ uncorrected, $p =$ not significant corrected). However, cross-system interactions were not significant, with the exception of negative functional interactions between the dmPFC and both right insula ($p = 0.002$, corrected) and left STG ($p < 0.001$, corrected). Most notably, there were no significant interactions between the key regions of ACC and dmPFC that correlated with specific appraisals and exhibited connectivity with vmPFC.

Second, we looked for correlations in the strength of the representation of the dominant appraisals in each region (e.g., emotional for ACC, utilitarian for dmPFC), for each pair of regions during the moral judgment task. This analysis allowed us to ask whether individuals who represent utilitarian appraisals more strongly in regions, such as the dmPFC or TPJ, represent emotional appraisals more weakly in the ACC or insula, as would be expected in the presence of inhibitory competition between the two systems. As shown in Figure 7*B*, results mirrored the patterns observed in regional connectivity: representations within the emotional appraisal network tended to positively correlate with each other (insula-ACC, $r_{(26)} = 0.47$, $p = 0.01$ uncorrected, $p =$ not significant, Bonferroni corrected, insula-STG, $r_{(26)} = 0.55$, $p = 0.03$ corrected), as did representations within

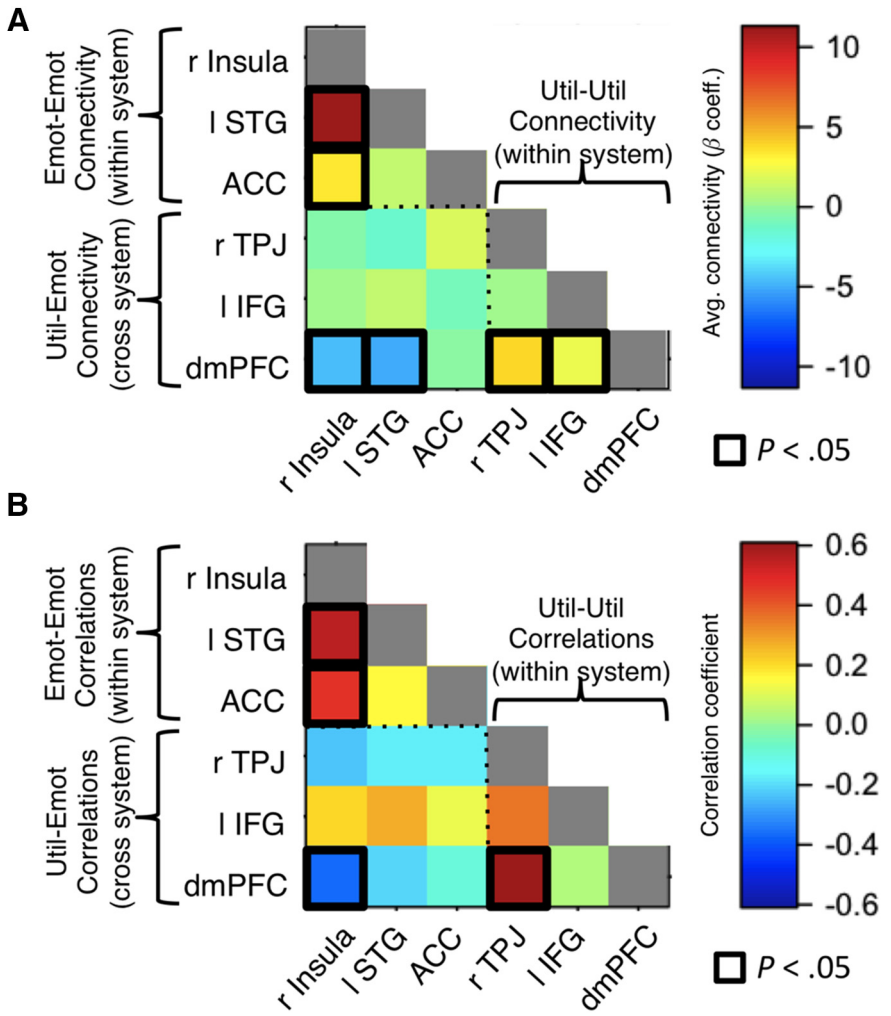

Figure 7. Interaction within and between the emotional and utilitarian appraisal networks during the moral judgment task. ***A***, Pairwise functional connectivity between regions, determined using FIR analyses and averaged over the time-points corresponding to the evaluation period. Colors represent the average regression coefficient across subjects. Black represents significant connectivity ($p < 0.05$, uncorrected for multiple comparisons). ***B***, Pairwise correlations between regions in the strength of the representation of the dominant attribute, computed for each individual using the estimated coefficients from GLM 2. Colors represent correlation coefficients. Black represents significant correlations ($p < 0.05$, uncorrected).

the utilitarian appraisal network (TPJ-dmPFC, $r_{(26)} = 0.61$, $p = 0.008$ corrected). However, cross-system representations did not correlate significantly, with the exception of a marginal negative correlation between utilitarian appraisals in the dmPFC and emotional appraisals in the insula ($r_{(26)} = -0.37$, $p = 0.05$ uncorrected, $p =$ not significant, corrected). These results provide additional support for the hypothesis that emotional and utilitarian appraisals are computed largely independently. We emphasize that the nature of the analyses cannot rule out the existence of some inhibitory interactions but suggests that their magnitude is likely to be small.

Discussion

Although considerable evidence suggests that moral judgments are influenced by emotional appraisals and utilitarian considerations (Suter and Hertwig, 2011; Amit and Greene, 2012; Conway and Gawronski, 2013), the neurocomputational basis of these signals, and how they interact to produce moral judgments, remains unclear. Our results support a relatively simple characterization of these processes. First, we found that anterior cingulate, insula, and superior temporal gyrus represent emotional appraisals, whereas temporoparietal junction and dorsomedial prefrontal cortex reflect utilitarian appraisals. Second, we found overall moral values represented in a separate region of vmPFC, similar but somewhat anterior to areas that encode decision values during simple economic choices (Bartra et al., 2013; Clithero and Rangel, 2014). Third, the pattern of responses within these regions, and their functional connectivity, supports the hypothesis that emotional and utilitarian appraisals are computed largely independently, and passed to the vmPFC to be integrated into an overall moral value.

Separate neural systems compute emotional and utilitarian appraisals

Previous work in this area has typically focused on differences in average activation between different types of moral dilemmas, such as those with or without negative, emotionally evocative features (Greene et al., 2001), having more or less conflict (FeldmanHall et al., 2014), or resulting in more or less utilitarian choices (Greene et al., 2004). These studies have provided invaluable insight into the dimensions shaping moral judgment and have established the involvement of regions, such as amygdala, TPJ, lateral prefrontal cortex, and vmPFC. However, they were not designed to identify the specific computations performed in these regions or to study how information flows between them. One notable exception is work by Shenhav and Greene (2014). This study tied amygdala response to overall negative emotional appraisals during moral judgment but did not report any regions in which responses correlated parametrically with utilitarian appraisals. Moreover, the study used complex tradeoff scenarios involving choices between two options, each with multiple components (e.g., push a man to his death and save five lives vs spare him and allow others to die) and did not collect subject-specific information on the different aspects of every option under consideration, which makes it impossible to address the questions of our study.

In contrast, our experiment was designed to identify separately regions encoding emotional appraisals, utilitarian appraisals, overall moral values, and the interactions among them. Our results suggest caution in interpreting previous findings. Although average amygdala response in our study was higher when rating emotional appraisals, this response did not correlate with idiosyncratic emotional ratings to standalone acts, which instead were encoded in areas such as insula, STG, and ACC. Previous work has shown stronger dorsolateral prefrontal cortex when the utilitarian option is favored (Greene et al., 2004). Although we had thus hypothesized that the dorsolateral prefrontal cortex might therefore represent utilitarian appraisals, we found more reliably encoding in re-

gions implicated in social cognition and Theory of Mind, such as the TPJ and dmPFC (Van Overwalle, 2009; Bzdok et al., 2012). To be clear, we do not view our findings as a definite refutation of previous interpretations of the computations performed in areas such as amygdala or dorsolateral prefrontal cortex, because multiple differences between the experimental paradigms make direct comparison difficult. However, our results underscore the need for further study in determining both when regions like these come on line during moral judgment, as well as the specific computations they perform.

Intriguingly, we also found that vlPFC and right STG responses correlated with utilitarian or emotional appraisals only during explicit rating tasks, whereas other regions showed consistent correlations even during the moral judgment task. Moreover, only ACC and dmPFC showed connectivity profiles consistent with passing information to the vmPFC for integration into an overall moral value. It is natural to speculate that ACC and dmPFC may themselves integrate lower-level features, perhaps represented in regions, such as amygdala, insula, STG, and TPJ, and could serve as relay stations that transform information into content usable by vmPFC.

Our results are consistent with previous work tying the ACC to emotional experience (Phan et al., 2002; Kober et al., 2008), and linking dmPFC and TPJ to moral judgment and social cognition (Bzdok et al., 2012). However, the role of regions like the mid-insula or STG in emotional representation remains less clear in the literature. Previous work sometimes associates mid-insula with pain processing (Wager et al., 2013), which could inform emotional appraisals invoked in our experiment. However, neither the mid-insula nor STG regions fall within canonical emotion circuits. Future work will thus need to determine the precise computations represented in these regions, including the representation of attributes beyond emotional and utilitarian appraisals. For example, real-world moral choices with selfish benefits often differ from hypothetical judgments (FeldmanHall et al., 2012), which suggests that additional attributes are informing judgment.

Implications for models of moral judgment

Our results support a nuanced version of the dual-systems view of moral choice (Greene et al., 2001; Haidt, 2001; Greene, 2005; Cushman et al., 2010; Conway and Gawronski, 2013; Cushman, 2013). Consistent with prior work, we find evidence that moral judgments are informed by emotional and utilitarian appraisals and that these appraisals are computed in nonoverlapping systems. However, our results do not support a model in which these systems directly inhibit each other in winner-take-all competition because we found only a negligible interaction effect in behavior, and no significant interaction effects in neural representations. Instead, our results support a simple attribute integration model, in which emotional and utilitarian appraisals are computed in distinct systems, which we here identify with ACC and dmPFC, respectively. These appraisals are integrated into an overall value judgment in a separate region of vmPFC.

The simple attribute integration model predicts that moral judgments are likely influenced by multiple attributes in parallel and suggests an important open question: what determines the relative weight that the attributes receive in the decision? Although utilitarian and emotional appraisals influenced judgments equally in the current context, it is natural to hypothesize that emotional appraisals might be more easily represented and weighted in other contexts because introspection and previous research suggest that they are harder to suppress than more deliberative utilitarian attributes.

Our results also shed light on models of the relationship between emotion and cognition in moral judgment. For example, in some theories, deliberative utilitarian considerations are simply *post hoc* justifications for emotional intuitions (Haidt et al., 1993; Haidt, 2001; Wheatley and Haidt, 2005; Schnall et al., 2008). Under this view, we might expect emotional appraisal regions, such as ACC and insula, to also be associated with the ratings in the utilitarian appraisal task. Yet, despite sizable correlation between the two appraisals, we found no overlap in the areas associated with them.

The neural bases of moral judgment

Our results suggest that complex moral decisions operate on many of the same principles at work in simple economic decisions (Fehr and Rangel, 2011; Rangel and Clithero, 2014), albeit with some important differences. Most studies of simple choice find overall-value signals in a region of the cingulate cortex closer to where we observed emotional appraisal representations (Kable and Glimcher, 2007; Bartra et al., 2013; Clithero and Rangel, 2014). Integrated moral value signals fell in a more anterior area of vmPFC that has been associated with the computation of more abstract values (Chib et al., 2009; McNamee et al., 2013; Clithero and Rangel, 2014). Although the location of this area is consistent with other studies of moral decision-making (Bzdok et al., 2012), it suggests a gradient of response in the vmPFC.

We also find that utilitarian appraisals are encoded negatively in dmPFC and TPJ (i.e., greater activity for less desirable values). Although we were surprised by this result, it is consistent with related work showing stronger TPJ responses to bad compared with good actions (Yoder and Decety, 2014). We emphasize that there is no a priori reason why the utilitarian appraisal code must be positive because measures of social cost (negative code) and social benefit (positive code) are equally useful in moral judgments. The significance and robustness of the negative code for utilitarian appraisals, and how this sign is flipped in overall value, represent an important question for future research.

Temporal dynamics of moral judgment

Despite the relatively low temporal resolution of fMRI (but for evidence of greater temporal sensitivity than previously assumed, see Katwal et al., 2012), we found changes in the signals represented in vmPFC over the course of a decision. For example, signals related to emotional and utilitarian appraisals appeared in this region at different times, and before representations of overall moral value. However, we observed little evidence that the earlier appearance of emotional appraisal signals resulted in a stronger average influence of emotional appraisals on choice behavior, perhaps because participants had sufficient time to integrate all the information. Time pressure influences decisions in a wide variety of paradigms, including nonsocial, social but nonmoral, and moral tasks (Ben Zur and Breznitz, 1981; Suter and Hertwig, 2011; Rand et al., 2012). Future work should investigate how time pressure affects the computations identified here and how this changes moral judgment.

References

Amit E, Greene JD (2012) You see, the ends don't justify the means: visual imagery and moral judgment. Psychol Sci 23:861–868. CrossRef Medline

Bartra O, McGuire JT, Kable JW (2013) The valuation system: a coordinate-based meta-analysis of bold fMRI experiments examining neural correlates of subjective value. Neuroimage 76:412–427. CrossRef Medline

Ben Zur H, Breznitz SJ (1981) The effect of time pressure on risky choice behavior. Acta Psychol 47:89–104. CrossRef

Boorman ED, Behrens TE, Woolrich MW, Rushworth MF (2009) How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron 62:733–743. CrossRef Medline

Bzdok D, Schilbach L, Vogeley K, Schneider K, Laird AR, Langner R, Eickhoff SB (2012) Parsing the neural correlates of moral cognition: Ale meta-analysis on morality, theory of mind, and empathy. Brain Struct Funct 217:783–796. CrossRef Medline

Chau BK, Kolling N, Hunt LT, Walton ME, Rushworth MF (2014) A neural mechanism underlying failure of optimal choice with multiple alternatives. Nat Neurosci 17:463–470. CrossRef Medline

Chib VS, Rangel A, Shimojo S, O'Doherty JP (2009) Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J Neurosci 29:12315–12320. CrossRef Medline

Clithero JA, Rangel A (2014) Informatic parcellation of the network involved in the computation of subjective value. Soc Cogn Affect Neurosci 9:1289–1302. CrossRef Medline

Conway P, Gawronski B (2013) Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. J Pers Soc Psychol 104:216–235. CrossRef Medline

Cushman F (2013) Action, outcome, and value: a dual-system framework for morality. Pers Soc Psychol Rev 17:273–292. CrossRef Medline

Cushman F, Young L, Greene JD (2010) Our multi-system moral psychology: towards a consensus view. In: The moral psychology handbook (Doris JM, ed), pp 47–71. Oxford: Oxford UP.

Decety J, Jackson PL (2006) A social-neuroscience perspective on empathy. Curr Dir Psychol Sci 15:54–58. CrossRef

Fehr E, Rangel A (2011) Neuroeconomic foundations of economic choice: recent advances. J Econ Perspect 25:3–30. CrossRef Medline

FeldmanHall O, Mobbs D, Evans D, Hiscox L, Navrady L, Dalgleish T (2012) What we say and what we do: the relationship between real and hypothetical moral choices. Cognition 123:434–441. CrossRef Medline

FeldmanHall O, Mobbs D, Dalgleish T (2014) Deconstructing the brain's moral network: dissociable functionality between the temporoparietal junction and ventro-medial prefrontal cortex. Soc Cogn Affect Neurosci 9:297–306. CrossRef Medline

Gallagher HL, Frith CD (2003) Functional imaging of 'theory of mind.' Trends Cogn Sci 7:77–83. Medline

Greene J (2005) Emotion and cognition in moral judgment: evidence from neuroimaging. Res Per Neurosci 57–66.

Greene J, Haidt J (2002) How (and where) does moral judgment work? Trends Cogn Sci 6:517–523. CrossRef Medline

Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) An fMRI investigation of emotional engagement in moral judgment. Science 293:2105–2108. CrossRef Medline

Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD (2004) The neural bases of cognitive conflict and control in moral judgment. Neuron 44:389–400. CrossRef Medline

Greene JD, Morelli SA, Lowenberg K, Nystrom LE, Cohen JD (2008) Cognitive load selectively interferes with utilitarian moral judgment. Cognition 107:1144–1154. CrossRef Medline

Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108:814–834. CrossRef Medline

Haidt J, Koller SH, Dias MG (1993) Affect, culture, and morality, or is it wrong to eat your dog. J Pers Soc Psychol 65:613–628. CrossRef Medline

Hutcherson CA, Plassmann H, Gross JJ, Rangel A (2012) Cognitive regulation during decision making shifts behavioral control between ventromedial and dorsolateral prefrontal value systems. J Neurosci 32:13543–13554. CrossRef Medline

Kable JW, Glimcher PW (2007) The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10:1625–1633. CrossRef Medline

Katwal SB, Gore JC, Gatenby JC, Rogers BP (2012) Measuring relative timings of brain activities using fMRI. Neuroimage 66C:436–448. CrossRef Medline

Kober H, Barrett LF, Joseph J, Bliss-Moreau E, Lindquist K, Wager TD (2008) Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. Neuroimage 42:998–1031. CrossRef Medline

Lim SL, O'Doherty JP, Rangel A (2013) Stimulus value signals in ventromedial PFC reflect the integration of attribute value signals computed in fusiform gyrus and posterior superior temporal gyrus. J Neurosci 33:8729–8741. CrossRef Medline

McNamee D, Rangel A, O'Doherty JP (2013) Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. Nat Neurosci 16:479–485. CrossRef Medline

Moll J, Zahn R, de Oliveira-Souza R, Krueger F, Grafman J (2005) Opinion: the neural basis of human moral cognition. Nat Rev Neurosci 6:799–809. CrossRef Medline

Phan KL, Wager T, Taylor SF, Liberzon I (2002) Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in pet and fMRI. Neuroimage 16:331–348. CrossRef Medline

Rand DG, Greene JD, Nowak MA (2012) Spontaneous giving and calculated greed. Nature 489:427–430. CrossRef Medline

Rangel A, Clithero J (2013) The computation of stimulus values in simple choice. In: Neuroeconomics: decision making and the brain (Glimcher PW, ed), pp 125–147. San Diego: Academic.

Saxe R, Powell LJ (2006) It's the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci 17:692–699. CrossRef Medline

Schnall S, Haidt J, Clore GL, Jordan AH (2008) Disgust as embodied moral judgment. Pers Soc Psychol B 34:1096–1109. CrossRef Medline

Shenhav A, Greene JD (2010) Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron 67:667–677. CrossRef Medline

Shenhav A, Greene JD (2014) Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. J Neurosci 34:4741–4749. CrossRef Medline

Suter RS, Hertwig R (2011) Time and moral judgment. Cognition 119:454–458. CrossRef Medline

Van Overwalle F (2009) Social cognition and the brain: a meta-analysis. Hum Brain Mapp 30:829–858. CrossRef Medline

Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E (2013) An fMRI-based neurologic signature of physical pain. N Engl J Med 368:1388–1397. CrossRef Medline

Wheatley T, Haidt J (2005) Hypnotic disgust makes moral judgments more severe. Psychol Sci 16:780–784. CrossRef Medline

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC (1996) A unified statistical approach for determining significant signals in images of cerebral activation. Hum Brain Mapp 4:58–73. CrossRef Medline

Yoder KJ, Decety J (2014) The good, the bad, and the just: justice sensitivity predicts neural response during moral evaluation of actions performed by others. J Neurosci 34:4161–4166. CrossRef Medline