Behavioral/Cognitive

Integrative Moral Judgment: Dissociating the Roles of the Amygdala and Ventromedial Prefrontal Cortex

Amitai Shenhav[1] and **Joshua D. Greene**[2]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08540, and [2]Department of Psychology, Harvard University, Cambridge, Massachusetts 02138

A decade's research highlights a critical dissociation between automatic and controlled influences on moral judgment, which is subserved by distinct neural structures. Specifically, negative automatic emotional responses to prototypically harmful actions (e.g., pushing someone off of a footbridge) compete with controlled responses favoring the best consequences (e.g., saving five lives instead of one). It is unknown how such competitions are resolved to yield "all things considered" judgments. Here, we examine such integrative moral judgments. Drawing on insights from research on self-interested, value-based decision-making in humans and animals, we test a theory concerning the respective contributions of the amygdala and ventromedial prefrontal cortex (vmPFC) to moral judgment. Participants undergoing fMRI responded to moral dilemmas, separately evaluating options for their utility (Which does the most good?), emotional aversiveness (Which feels worse?), and overall moral acceptability. Behavioral data indicate that emotional aversiveness and utility jointly predict "all things considered" integrative judgments. Amygdala response tracks the emotional aversiveness of harmful utilitarian actions and overall disapproval of such actions. During such integrative moral judgments, the vmPFC is preferentially engaged relative to utilitarian and emotional assessments. Amygdala-vmPFC connectivity varies with the role played by emotional input in the task, being the lowest for pure utilitarian assessments and the highest for pure emotional assessments. These findings, which parallel those of research on self-interested economic decision-making, support the hypothesis that the amygdala provides an affective assessment of the action in question, whereas the vmPFC integrates that signal with a utilitarian assessment of expected outcomes to yield "all things considered" moral judgments.

Key words: decision making; emotion; reasoning; fMRI; amygdala; vmPFC

Introduction

Moral dilemmas are both useful and interesting because they evoke competing, incompatible judgments, revealing the fault lines in moral cognition (Cushman and Greene, 2012). A decade's research highlights a critical dissociation between automatic and controlled influences on moral judgment, which is subserved by distinct neural structures (Greene et al., 2001; Greene et al., 2004; Mendez et al., 2005; Schaich Borg et al., 2006; Ciaramelli et al., 2007; Koenigs et al., 2007). Although it is clear that automatic and controlled processes sometimes compete in the production of moral judgments (Greene et al., 2004, 2008; Cushman et al., 2012; Suter and Hertwig, 2011; Paxton et al., 2012; Conway and Gawronski, 2013), it is not known how such competitions are resolved to yield "all things considered" judgments. Here, we examine such integrative moral judgments, test-

ing a theory concerning the respective contributions of the amygdala and ventromedial prefrontal cortex (vmPFC).

The need for integration is illustrated by the classic "footbridge" dilemma (Foot, 1978; Thomson, 1986) in which one can save five lives at the cost of one by pushing someone off of a footbridge and into the path of a runaway trolley. The utilitarian response—favoring pushing because it saves more lives—is preferentially supported by controlled cognitive processes (Greene et al., 2008; Suter and Hertwig, 2011; Paxton et al., 2012). The alternative nonutilitarian, or characteristically "deontological," response is preferentially supported by automatic emotional responses (Greene et al., 2008; Suter and Hertwig, 2011; Paxton et al., 2012) that depend in some way on the vmPFC and, perhaps, the amygdala. Several studies show that patients with vmPFC damage make more utilitarian judgments (Ciaramelli et al., 2007; Koenigs et al., 2007; Moretto et al., 2010; Thomas et al., 2011). However, it is not clear whether the vmPFC is responsible for generating anti-utilitarian affective responses, if its role is to integrate such responses into judgments, or both. The evidence implicating the amygdala is less extensive. Subregions of the amygdala respond particularly strongly to "personal" dilemmas such as the footbridge case (Greene et al., 2004) and less so in individuals with psychopathic traits (Glenn et al., 2009). Although the amygdala has been implicated in moral cognition more generally (King et al., 2006; Berthoz et al., 2006; Schaich
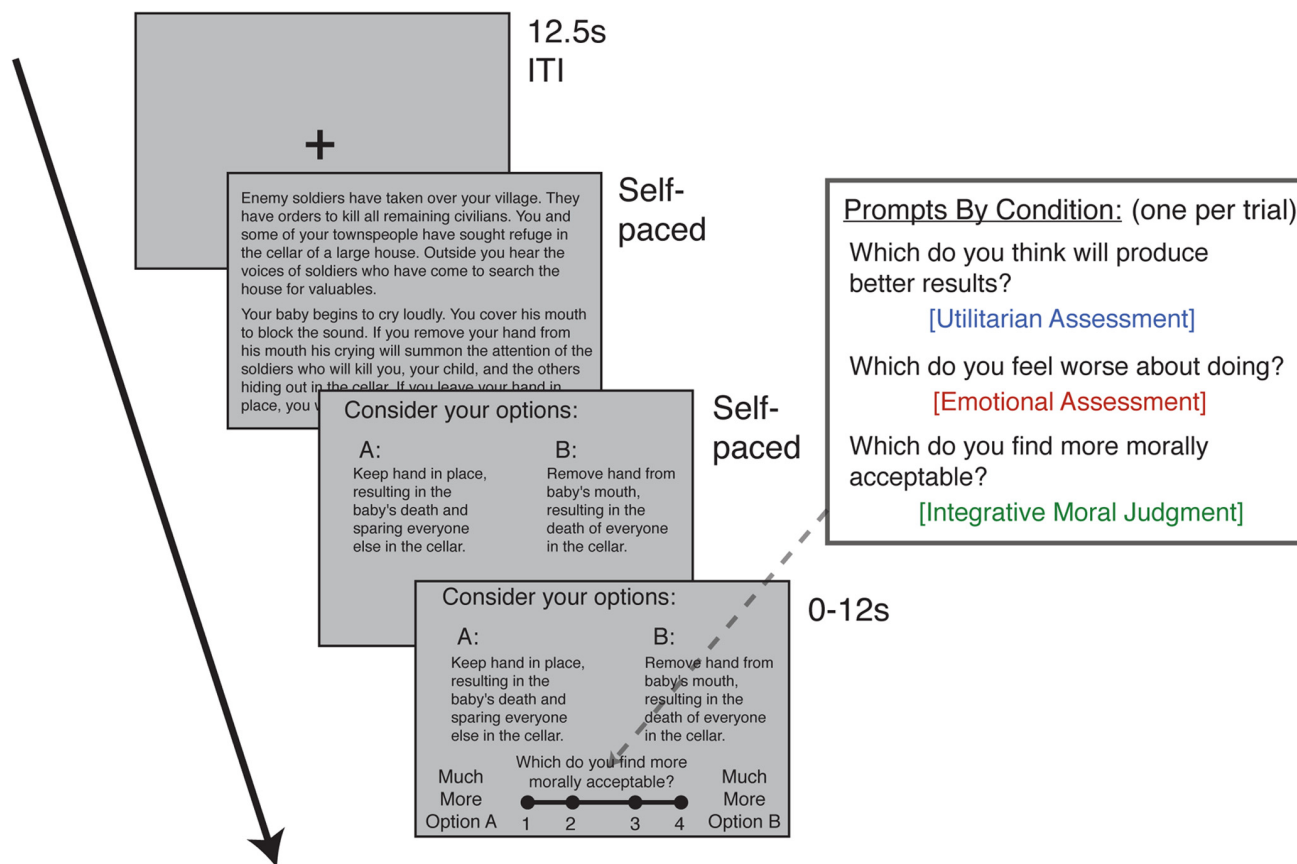
Figure 1. Task design. The first slide describes the dilemma, including two possible options. The utilitarian option minimizes the overall amount of harm (e.g., saves five lives), but involves actively harming someone. The nonutilitarian/deontological option does not involve active harm, but it fails to minimize harm. The second screen summarizes the two options, randomly labeling them "A" or "B." The third screen prompts the subject to respond in one of three ways depending on condition. Subjects have 12 s to respond using a four-point scale.

Borg et al., 2011), its activity has not been connected to specific responses to moral dilemmas.

Here, we test a neuroscientific theory of integrative moral judgment suggested by studies of nonmoral decision-making in humans and animals (Wallis, 2007; Rangel and Hare, 2010; Schoenbaum and Esber, 2010; Grabenhorst and Rolls, 2011; Padoa-Schioppa and Cai, 2011) in conjunction with research on moral judgment (Blair, 2007; Shenhav and Greene, 2010). We hypothesize that the amygdala enables automatic emotional responses to "personally" harmful actions (Greene et al., 2009), whereas the vmPFC integrates such responses into "all things considered" judgments, weighing them against utilitarian considerations. We test this hypothesis by comparing and examining three related judgment tasks: simple emotional assessments (Which option feels worse?), simple utilitarian assessments (Which will produce better results?), and integrative, "all things considered" judgments (Which is more morally acceptable?).

Materials and Methods

Subjects. A total of 39 healthy, right-handed subjects (20 female) with no reported history of neurological or affective disorders were recruited for the fMRI experiment. Of these, four were excluded before fMRI analysis, one due to technical difficulties, one for excessive excluded trials (see Behavioral analysis, below), one for later reporting having misconstrued the instructions, and one for falling asleep during the task. After the exclusions, data from 35 subjects (19 female, mean age 22.9 years, range 18–32) were analyzed. In addition, one of the six task blocks was excluded from analysis a priori for two included subjects due to technical difficulties.

Image acquisition. Images were acquired using a Siemens 3T Treo MR magnet and a 12-channel RF head coil. We acquired functional image volumes as T2*-weighted echo-planar images (EPIs) with the following parameters: 41 interleaved slices, 2500 ms TR, 28 ms TE, 2.5 mm thickness, 0.75 mm gap, 64 × 64 matrix, 200 mm FOV (resulting in an inplane voxel size of 3.125 × 3.125 × 2.5 mm). Our fMRI sequence and slice prescription were optimized for reducing signal loss/distortion in the orbitofrontal cortex (based on recommendations of Deichmann et al., 2003; Weiskopf et al., 2006), including the use of a modified *z*-shim prepulse moment and 30° tilt of our slice prescription off the AC/PC line. Due to the dynamic and unpredictable length of individual trials (the stages within each trial were self-paced, with the exception of the judgment phase which was capped at 12 s), runs were of variable lengths. Each run also included an additional 10 s of fixation at the beginning (to allow for the fMRI signal to reach steady-state), and the corresponding four EPI volumes were discarded from further analysis. Each session included the acquisition of a high-resolution T1-weighted anatomical image (1 mm isotropic voxels), during which time subjects read instructions for the task (as a series of sequential screens) and performed three practice trials.

Task design. Participants underwent fMRI while responding to 48 "trolley"-type moral dilemma, ones in which maximizing the number of lives saved requires actively harming one or more individuals (full text available by request). During each trial, subjects proceeded through three screens describing the dilemma (Screen 1), summarizing the options (Screen 2), and requesting a response (Screen 3; Fig. 1). Screens 1 and 2 were identical across conditions. Screen 3 varied by condition, prompting the subject to compare the two options in one of three ways:

Utilitarian assessment (UA) condition: "Which [option] do you think will produce better results?"

Emotional assessment (EA) condition: "Which [option] do you feel worse about doing?"

Integrative moral judgment (IMJ) condition: "Which [option] do you find more morally acceptable?"

Subjects responded using a 1–4 scale anchored by the phrases "Much more Option A" and "Much more Option B." Dilemmas and conditions (prompt types) were randomly ordered. There were 16 dilemmas per condition. Randomization was performed independently for each subject such that dilemma content was decoupled from condition across subjects. Each session included six blocks of eight trials each.

Two critical design features warrant attention. First, the trial type (condition: IMJ, EA, or UA) was not revealed to subjects until the final stage of each trial. This enabled us to better isolate the effect of condition (assessment/judgment type). Second, the spatial location (left or right) and label ("Option A" or "Option B") of the utilitarian option was randomly varied. This decoupled the content of the judgment from the motor response and linguistic representation associated with the choice.

Behavioral analysis. We analyzed the behavioral data (judgment ratings and response times) using mixed-effects analyses, modeling intercepts and slopes for each participant as random effects. To determine whether and to what extent the "all things considered" judgments made in the IMJ condition were influenced by emotional and utilitarian considerations, we also compared these judgments to the average EAs and UAs made by other participants to the same scenario (Fig. 2). Individual participant β estimates were obtained in a nearly identical fashion by performing regressions of IMJ ratings on normed EA and UA ratings for each participant separately. The meanings of the numerical ratings given in each trial varied with the arbitrary label assignments (A or B; see previous paragraph). We therefore recoded the ratings such that all ratings apply to the utilitarian option, indicating the degree to which the utilitarian option feels worse, produces better results, or is more morally acceptable. To omit trials in which the subject was likely inattentive, we excluded trials in which RTs for reading the context, behavioral options, and judgment were respectively <4 s, 2 s, and/or 0.5 s. We also excluded trials that timed out before the subject submitted a response in the final stage (12 s time window). An average of 2.6 (median = 2) trials out of 48 (5%) were excluded per participant.

fMRI analysis. fMRI analysis was performed in SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London). Data preprocessing included motion correction, slice timing correction, normalization of both functional volumes and the high-resolution anatomical volume to standardized (MNI) templates (including resampling to 2 mm isotropic voxels), and spatial smoothing with a 6 mm Gaussian kernel.

Further analysis of fMRI time courses focused on events ("stick" functions with 0 s duration) modeled in the judgment phase of each trial, with an onset 750 ms after the presentation of the judgment prompt (to allow for reading/encoding). These events were convolved with a canonical hemodynamic response function. We performed two sets of whole-brain fMRI analyses, one aimed at identifying effects of condition and one aimed at identifying effects of behavior within condition. Both used a general linear model (GLM) with analysis-specific controls for RT.

The GLM analyses aimed at identifying effects of condition used serially orthogonalized parametric regressors which modeled average differences between conditions after first modeling the average signal change across all trials and RT differences across all conditions. Specifically, this GLM modeled a single indicator function across all judgments. This single event type was then parametrically modulated by the following parameters (in the order given): (1) judgment RT; (2) a parameter with 1 for UA trials, −1 for IMJ trials, and 0 otherwise; and (3) a parameter with 1 for IMJ trials, −1 for EA trials, and 0 otherwise. Contrasts for IMJ versus UA and for IMJ versus EA were accomplished by performing a *t* test over the second and third parameters, respectively. Contrasts for EA versus UA were accomplished with a joint *t* test over the second and third parameters. This approach is recommended when performing categorical (e.g., condition-wise) contrasts subsequent to controlling hierarchically for a continuous parameter (e.g., RT; R. Henson, personal communication, 2009).

Additional GLM analyses aimed to identify effects of behavior within condition. These analyses modeled each condition as a separate event type (in contrast to the GLM described in the previous paragraph, which modeled a single event type for the judgment period). Each condition was then separately parametrically modulated by (1) judgment RT and (2) rating. For the UA condition, the rating regressor was excluded due to minimal variability within and across subjects. In a follow-up analysis, a separate GLM included a binary regressor that differentiated between trials in which the utilitarian option was identified as producing "much better" versus only "somewhat better" results. These analyses excluded three participants who gave only one of these two responses.

All analyses included an independent regressor modeling activity associated with the reading of the text before the judgment prompt; that is, the point at which the trial's condition is revealed. This was done to control for variability unrelated to condition. Activity associated with Screens 1–2 (descriptions of the dilemma context and possible actions) was jointly modeled as related to a single epoch of variable duration. Finally, regressors were included to model the mean BOLD signal and linear trend across each block. First-level (single subject) contrasts were performed over individual parametric regressors of interest. Second-level (group) random effects analyses used one-sample *t* tests over the first-level parametric maps. Based on our a priori region-specific hypotheses, FWE small-volume correction was performed using anatomically defined masks in the amygdala and vmPFC. For completeness, we also report any nonsignificant clusters that appear in these ROIs with a voxelwise-uncorrected threshold of $p < 0.05$. ROIs were generated using the Automated Anatomical Labeling (AAL) toolbox in SPM8 (Tzourio-Mazoyer et al., 2002). Left and right amygdalae were tested separately and the bilateral vmPFC ROI was generated by combining regions labeled as the medial orbital frontal component of the superior frontal gyrus and the gyrus rectus. For visualization purposes, cortical and subcortical activations are shown on normalized volumes thresholded at voxelwise $p < 0.005$, extent-thresholded at 40 and 10 voxels, respectively. *Post hoc* analyses were performed to confirm that positive findings in the amygdala survive correction with a single bilateral ROI and to explore whether vmPFC activations varied when correcting separately for left and right hemisphere ROIs.

We performed a second type of analysis, adapted from previous studies of financial decision-making (Kuhnen and Knutson, 2005; Knutson et al., 2007), in which BOLD reactivity in one or more regions was used to predict behavioral ratings, rather than the reverse. Specifically, we obtained trialwise estimates of BOLD activity in our a priori ROIs and included these estimates in within-subject regressions predicting behavioral ratings, as well as additional analyses meant to examine changes in interregion correlations (functional connectivity) between conditions. For these analyses, we performed an additional whole-brain GLM that included a separate event regressor after the onset of each trial's judgment period (cf. Rissman et al., 2004) regardless of condition type, as well as a single variable-duration block regressor to model all prejudgment reading periods (as in the previous GLMs). This GLM also included the same regressors of no interest to model means and linear trends for each block. We then extracted averaged trialwise β estimates within each of our ROIs. We normalized (*z*-scored) these values within subject. To reduce the influence of outliers, we also Winsorized these β estimates at 3 SDs (i.e., the maximum absolute *z*-score was set to 3). We performed the following within-subject regressions, pooling regression coefficients to then obtain group-level statistics (as with our whole-brain GLMs).

We regressed behavioral ratings in the EA and IMJ conditions on RT and β estimates from amygdala and vmPFC. Primary analyses used amygdala β estimates from a bilateral ROI, but secondary analyses are included to describe differences when the ROI is left or right lateralized. These analyses use one-sided *t* tests when testing directional predictions generated by our theoretical model and tested accordingly by our whole-brain GLM contrasts. We also regressed vmPFC reactivity (β estimates) on RT, behavioral rating, condition (conditionwise regressors constructed as in whole-brain GLM 1), amygdala reactivity, and the interaction between condition and amygdala reactivity. This analysis is structurally similar to a psychophysiological interaction analysis (cf. Gitelman et al., 2003; see also Rissman et al., 2004). It allowed us to
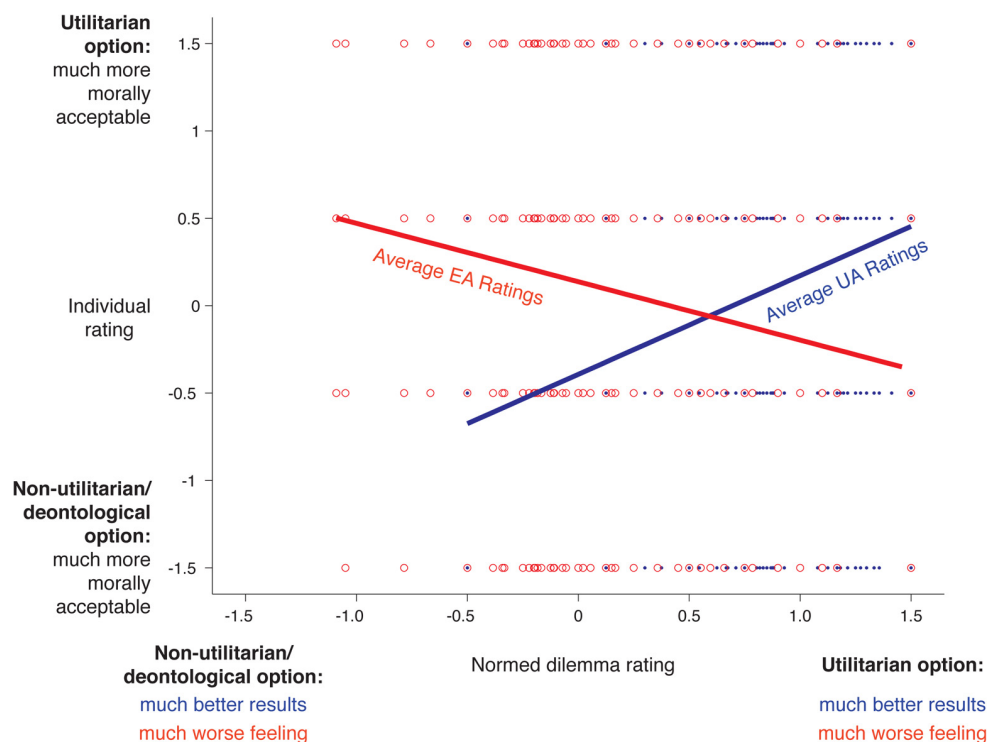
Figure 2. "All things considered" judgments (IMJ) reflect integration of competing moral considerations. Average regression slopes predicting "all things considered" moral judgments (IMJ) for each participant based on average EA and UA responses to the scenario from other participants. Participants rated the utilitarian option as more morally acceptable if other participants perceived that option as producing better results (high UA) and/or felt less bad about choosing that option (low EA). Scatterplot shows data from all individual IMJ trials relative to EA (red) and UA (blue) ratings.

compare changes in interregion correlation (functional connectivity) depending on condition.

Note that the whole-brain GLMs and the regressions described in the previous paragraph are complementary in at least two ways. First, they test predictions in opposite directions (predicting behavioral ratings from the activity of multiple brain regions versus the reverse). Second, whereas the latter regressions rely on signal averaged over an ROI, the whole-brain GLMs allow us to identify significant clusters within the same ROIs without requiring that the average signal across all voxels in the ROI be significantly modulated for a given contrast.

Results

Behavioral results

As expected, in the UA condition ("Which [option] do you think will produce better results?") subjects consistently rated the utilitarian option as producing better results ($M_{UA} = 0.91$, $SD_{UA} = 0.38$; scale ranging from -1.5 to 1.5 with ratings at the positive end of the scale favoring the utilitarian option). Here, ratings typically varied within the upper end of the scale, distinguishing between options perceived as moderately versus highly utilitarian. In the EA condition ("Which [option] do you feel worse about doing?") and in the IMJ condition ("Which [option] do you find more morally acceptable?"), subjects tended to use the entire scale, with mean ratings near the middle of the scale ($M_{EA} = 0.01$, $SD_{EA} = 0.46$; $M_{IMJ} = 0.06$, $SD_{IMJ} = 0.51$). UA responses were significantly faster than IMJ responses ($M_{UA} = 4.1$ s, $M_{IMJ} = 4.5$ s; $F_{(1,65.2)} = 6.82$, $p < 0.02$), which were in turn significantly faster than EA responses ($M_{EA} = 5.1$ s; $F_{(1,66.0)} = 20.5$, $p < 10^{-4}$; UA vs EA: $F_{(1,65.9)} = 50.9$, $p < 10^{-9}$).

Finally, we examined the relationship between a participant's IMJ ratings for each scenario and the average EA and UA ratings given by other participants in response to the same dilemmas.

This tested the hypothesis that IMJ ratings are a function of the relative emotional and/or utilitarian weights attached to the options under consideration. With both EA and UA ratings entered simultaneously into a mixed-effects multiple regression, we found that both were highly significant predictors of IMJ (moral acceptability) ratings and in opposite directions ($b_{UA} = 0.60$, $SE_{UA} = 0.08$, $t_{UA}(42.4) = 7.7$, $p < 10^{-8}$; $b_{EA} = -0.30$, $SE_{EA} = 0.05$, $t_{EA}(35.1) = -5.7$, $p < 10^{-5}$; Fig. 2). This is consistent with our hypothesis that "all things considered" judgments involve the integration of competing valuations based respectively on utilitarian assessments and emotional responses. Across dilemmas ($N = 48$), we observed significant correlations between each pair of normed ratings ($r_{EA,UA} = -0.42$, $r_{IMJ,UA} = 0.52$, $r_{IMJ,EA} = -0.60$, $ps < 0.005$).

Neuroimaging results

vmPFC and integrative moral judgment

To identify neural activity specifically associated with integrative, "all things considered" judgments, we compared the IMJ condition with the UA and EA conditions. The separate contrasts IMJ > UA and IMJ > EA both revealed increased activity in vmPFC (peak MNI coordinates [x, y, z] and small-volume corrected p-value, IMJ > UA: $-6, 26, -16$, $p < 0.05$; IMJ > EA: 0, $36, -12$, $p < 0.01$; Fig. 3A,B, Table 1). Likewise, the conjunctive contrast IMJ>UA ∩ IMJ>EA revealed increased activity for integrative moral judgment in an overlapping region of vmPFC (Fig. 3C). Post hoc analyses indicated that these effects were more reliable for the left vmPFC (IMJ > UA: $-6, 26, -16$, $p < 0.02$; IMJ > EA: $-2, 36, -12$, $p < 0.01$) than the right vmPFC (IMJ > UA: $2, 42, -14$, $p = 0.67$; IMJ > EA: $0, 36, -10$, $p < 0.01$).
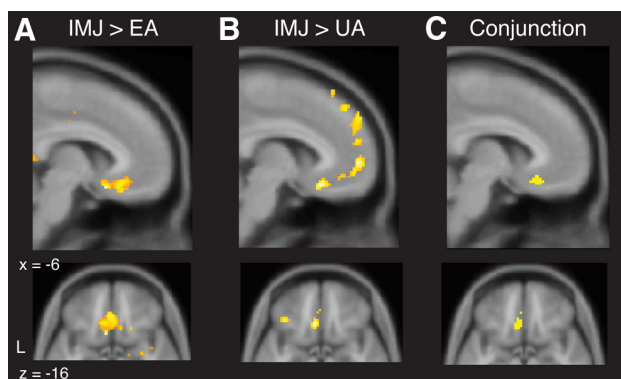
Figure 3. The vmPFC exhibits increased BOLD signal for IMJ relative to EA and UA. Shown are whole-brain results for IMJ > EA (**A**) and IMJ > UA (**B**). Analyses use an anatomically defined vmPFC ROI. For visualization, maps are thresholded at voxelwise $p < 0.005$ with an extent threshold of 40 voxels. **C**, The conjunction of these two contrasts.

Amygdala and emotional assessment

Our hypothesis that the amygdala enables automatic emotional evaluations of morally salient actions makes two predictions. First, in the EA condition, amygdala activity should correlate positively with negative emotional assessments of the utilitarian option (rating it as "feeling worse"). We predict a positive correlation, rather than a negative correlation or no correlation, based on prior research indicating that the strongest emotional influences in similar dilemmas are negative emotional reactions to harmful utilitarian actions (Greene et al., 2001, 2004; Koenigs et al., 2007; Glenn et al., 2009). Confirming this prediction, we observed parametrically increasing BOLD signal within the left amygdala for more negative ratings of how bad the utilitarian option feels (peak: −28, −4, −22, $p < 0.01$ SVC; Fig. 4A) In the right amygdala, we identified a cluster exhibiting a nonsignificant effect in the same direction (peak: 32, −4, −20, $p = 0.26$ SVC).

Our hypothesis concerning the amygdala's present functional role makes a second prediction. Insofar as the amygdala's response bears on moral judgment, increased amygdala activity should correlate negatively with utilitarian judgment in the IMJ condition. Confirming this prediction, we observed parametrically increasing BOLD signal within the right amygdala as subjects rated the utilitarian option as being less morally acceptable "all things considered" (peak: 18, −2, −16, $p < 0.05$ SVC; Fig. 4B). Here, too, we identified a cluster in the opposite hemisphere (left amygdala) exhibiting a nonsignificant effect in the same direction (peak: −24, 0, −12, $p = 0.17$ SVC). Both this right amygdala cluster for IMJ ratings and the left amygdala cluster identified for EA ratings survive correction with a bilateral amygdala mask (EA: −28, −4, −22, $p < 0.02$; IMJ: 18, −2, −16, $p < 0.05$).

Therefore, we found anti-utilitarian effects in the amygdala in two independent analyses and within two separate conditions (EA and IMJ). However, the first effect was observed in the left amygdala and the second in the right. This could be due to lateralized function or simply to subthreshold effects on one or both sides. To address this question we used a more liberal threshold ($p < 0.05$, uncorrected) within our a priori anatomically defined amygdala ROIs. This revealed effects of both EA and IMJ ratings in an overlapping region of the left amygdala, arguing against a strong lateralization interpretation and suggesting instead that these two effects, at least on the left, may reflect common processes (Fig. 4C). Notably, neither the EA ratings nor the IMJ ratings correlated with vmPFC signal in the direction seen for the

amygdala (EA: peak: −10, 44, −12, $p = 0.89$ SVC; IMJ: 0 voxels with uncorrected $p < 0.05$; but see the next section).

As noted earlier, EA and UA ratings were moderately negatively correlated. This raises the possibility that the amygdala signal tracks EA ratings, not because of its role in (negative) emotional assessment, but because it is involved in (positive) utilitarian assessment. To test the viability of this alternative interpretation, we investigated whether amygdala signal differentiates between UA trials in which participants rate the utilitarian option as producing outcomes that are "much better" versus "somewhat better" (as noted earlier, nearly all of the UA rating variance is confined to these two responses). We found no significant clusters in either amygdala ROI exhibiting increased activity with higher UA ratings (L peak: −22, −2, −20, $p > 0.50$ SVC; R peak: 32, 4, −26, $p > 0.25$ SVC). This is consistent with our interpretation of the amygdala-EA correlation as being related to negative emotional assessment, rather than positive utilitarian assessment.

Amygdala-vmPFC interactions in generating moral judgment

Our final set of analyses simultaneously examined the relative contributions of the amygdala and vmPFC to behavioral ratings. We extracted trial-by-trial β estimates from bilateral amygdala and vmPFC ROIs and entered these as predictors in a regression predicting ratings within a given condition, while controlling for RT. We found that amygdala again tracked ratings of greater emotional aversiveness (b = 0.094, SE = 0.064, $t_{(34)} = 1.47$, 1-tailed $p = 0.075$) and lower moral acceptability (b = −0.11, SE = 0.059, $t_{(34)} = -1.89$, 1-tailed $p = 0.033$) of the utilitarian option (see Materials and Methods for justification of 1-tailed tests). Although the former result is marginally significant, stronger effects emerge for EA ratings in an analysis restricted to the left amygdala (b = 0.18, SE = 0.082, $t_{(34)} = 2.22$, 1-tailed $p = 0.017$) and for IMJ ratings when the analysis is restricted to the right amygdala (b = −0.12, SE = 0.053, $t_{(34)} = -2.37$, 1-tailed $p = 0.012$), which is consistent with the results of our whole-brain analysis. Interestingly, while controlling for bilateral amygdala reactivity, we saw a significant relationship between vmPFC betas and judgments that the utilitarian option is more morally acceptable (b = 0.15, SE = 0.035, $t_{(34)} = 2.10$, 2-tailed $p = 0.044$). We observed a nonsignificant negative correlation between vmPFC betas and ratings of emotional aversiveness (b = −0.087, SE = 0.055, $t_{(34)} = -1.58$, 2-tailed $p = 0.124$). Performing the equivalent logistic regression on UA ratings again failed to identify a significant relationship between bilateral amygdala reactivity and more positive utilitarian assessments (b = −0.87, SE = 2.23, $t_{(31)} = -0.39$, 2-tailed $p > 0.65$).

Individuals varied in the extent to which their IMJ ratings were correlated with trial-to-trial fluctuations in amygdala signal. Individuals also varied in the extent to which their IMJ ratings were correlated with the average EA rating for a given dilemma. If the amygdala mediates the influence of emotional aversiveness on IMJ ratings and an individual's IMJ ratings were heavily influenced by emotional assessments, then both correlations should be strong. Likewise, both correlations should be weak in individuals whose judgments were relatively immune to emotional influence. Therefore, across participants, the strengths of these two correlations should themselves be correlated. With this in mind, we examined the relationship between (1) participant-level β estimates reflecting the relationship between IMJ ratings and bilateral amygdala signal (from the previous regression) and (2) the relationship between IMJ ratings and the average EA rating for a given dilemma (from a regression that also included average UA ratings). Consistent with our hypothesis, we found a marginally

Table 1. Whole-brain exploratory analyses for contrasts of interest

Side	Region	Cluster-level p-value	Cluster size (voxels)	Peak voxel Z-score	Peak voxel p-value	MNI coordinates (mm)
EA > IMJ						
B	Pre-SMA	<0.0001	1115	5.28	<0.0001	−4, 22, 62
B	Precuneus	0.0006	533	5.14	<0.0001	−8, −58, 44
L	Temporoparietal junction, middle temporal gyrus	<0.0001	2376	5.06	<0.0001	−50, −52, 26
L	Middle frontal gyrus (BA 8)	<0.0001	1322	4.91	<0.0001	−42, 20, 46
L	Ventrolateral PFC	0.0054	390	4.21	<0.0001	−50, 18, −6
R	Temporoparietal junction	0.0027	434	3.97	<0.0001	46, −58, 30
L	Inferior frontal gyrus	0.0028	432	3.54	0.0002	−42, 52, 2
R	Cerebellum*	0.0799	232	3.61	0.0002	22, −80, −32
IMJ > EA						
R	Primary, secondary somatosensory cortex, central insula, amygdala	<0.0001	3836	4.80	<0.0001	60, −28, 22
B	Medial orbitofrontal cortex/vmPFC	0.0003	576	4.58	<0.0001	0, 36, −12
L	Primary, secondary somatosensory cortex, central insula, amygdala	<0.0001	1183	4.27	<0.0001	−36, 6, −8
R	Postcentral gyrus*	0.0743	236	4.41	<0.0001	30, −46, 72
IMJ > UA						
B	Dorsomedial PFC*	0.0538	266	3.64	0.0001	0, 52, 38
UA > IMJ						
R	Inferior frontal gyrus (BA 9/44)	0.0331	295	4.66	<0.0001	48, 6, 30
L	Intraparietal sulcus	0.0074	389	4.01	<0.0001	−32, −72, 32
R	Intraparietal sulcus	0.0109	364	3.95	<0.0001	30, −74, 40
UA > EA						
R	Posterior cingulate cortex	0.0114	370	4.61	<0.0001	16, −36, 36
R	Primary, secondary somatosensory cortex, central insula, inferior frontal gyrus (BA 9/44)	<0.0001	2339	4.15	<0.0001	50, −36, 34
R	Intraparietal sulcus	<0.0001	947	4.10	<0.0001	30, −72, 40
L	Primary, secondary somatosensory cortex, central insula	0.0018	497	3.87	0.0001	−48, −4, 6
EA > UA						
L	Temporoparietal junction	<0.0001	955	4.62	<0.0001	−54, −60, 26
B	Precuneus	0.0084	390	4.56	<0.0001	−6, −58, 44
R	Cerebellum	0.017	344	4.39	<0.0001	22, −80, −32
L	Ventrolateral PFC	0.0001	713	4.29	<0.0001	−54, 18, 6
L	Middle frontal gyrus (BA 8)	0.0003	643	4.21	<0.0001	−32, 14, 42
B	Pre-SMA, superior frontal gyrus	<0.0001	1054	4.11	<0.0001	−4, 34, 62
L	Inferior frontal gyrus	0.0018	497	3.96	<0.0001	−46, 42, 0
R	Temporoparietal junction	0.0496	277	3.87	0.0001	58, −60, 24

Significant clusters of activation for each of the condition-wise contrasts between EA, UA, and IMJ thresholded at voxel-wise $p < 0.005$ and whole-brain FWE corrected at cluster-wise $p < 0.05$ are shown. Clusters that were marginally significant ($p < 0.10$) at this cluster-wise threshold are marked with an asterisk (*). No activations survived this whole-brain correction when examining either positive or negative correlations with EA or IMJ ratings.
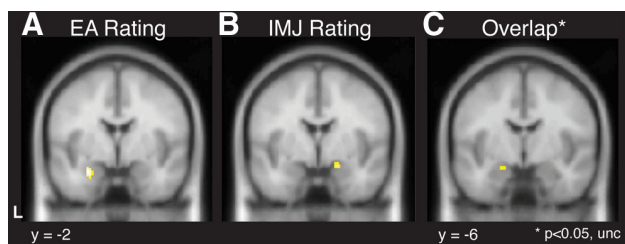

Figure 4. The amygdala's role in EA and IMJ. **A**, BOLD activity in the left amygdala tracked higher ratings of "feeling worse" about the utilitarian option in the EA condition. **B**, BOLD activity in the right amygdala tracked lower moral acceptability ratings for the utilitarian option in the "all things considered" (IMJ) condition. Analyses use anatomically defined amygdala ROIs. For visualization, maps are thresholded at voxelwise $p < 0.005$ with an extent threshold of 10 voxels. **C**, Overlapping activation for these two contrasts at a much reduced threshold ($p < 0.05$, uncorrected, masked by the anatomically defined bilateral amygdala ROI) suggests that these two processes may not be lateralized, despite the apparent lateralization in **A** and **B**.

significant correlation between these two sets of β estimates (robust regression coefficient = 0.28, SE = 0.14, 2-tailed $p = 0.066$).

These analyses allowed us to examine relationships in trial-to-trial reactivity between our key regions of interest (functional connectivity). Specifically, we asked whether bilateral amygdala and vmPFC β estimates were correlated and whether the strengths of these correlations varied by condition. Here, we controlled for main effects of condition, RT, and rating. We found

that vmPFC and amygdala β estimates were highly correlated (b = 0.42, SE = 0.041, $t_{(34)}$ = 10.2, 2-tailed $p < 10^{-11}$) and that, relative to the IMJ condition, the strength of this correlation increased during EA judgments (b = 0.15, SE = 0.036, $t_{(34)}$ = 4.0, 2-tailed $p = 0.0003$) and decreased during UA judgments (b = −0.082, SE = 0.041, $t_{(34)}$ = −1.77, 2-tailed $p = 0.05$). Therefore, the vmPFC and amygdala appear to be most tightly coupled when the task is to make an explicit emotional assessment and least tightly coupled when the task is to make an explicit utilitarian assessment. Coupling between amygdala and vmPFC is intermediary during "all things considered" judgment, which is consistent with the hypothesis that such judgments require the vmPFC to integrate information from both the amygdala and other brain regions responsible for utilitarian assessment.

Discussion

Here, we attempt to dissociate the roles of two neural regions believed to be critical for moral judgment: the vmPFC and the amygdala. Our analysis builds on insights achieved through research on the cognitive neuroscience of moral judgment (Greene et al., 2001, 2004; Mendez et al., 2005; Blair, 2007; Ciaramelli et al., 2007; Koenigs et al., 2007; Glenn et al., 2009; Harenski et al., 2010; Schaich Borg et al., 2011), as well as domain-general decision processes (Wallis, 2007; Rangel and Hare, 2010; Schoenbaum and Esber, 2010; Grabenhorst and Rolls, 2011; Padoa-Schioppa and Cai, 2011). We hypothesized that moral judgment depends on integrat-

ing emotional responses with utilitarian assessments. With this in mind, we had participants separately evaluate options for their utility, emotional aversiveness, and overall moral acceptability. We find that both utilitarian and emotional assessments are integrated into overall moral judgments and that the vmPFC is more active during these judgments than when making either of the component assessments. Thus, we provide evidence that the vmPFC plays a critical role in integrative moral judgment. We also provide evidence for the amygdala's role in the emotional assessment of morally salient options. More specifically, we show that increased amygdala response is associated with the evaluation of utilitarian options as more emotionally aversive and less morally acceptable. Likewise, we find a (marginally significant) correlation across individuals between reliance on emotional information and the extent to which amygdala activity is correlated with judgment on a trial-by-trial basis. Finally, we see the highest degree of coupling between amygdala and vmPFC when participants are asked to make purely emotional assessments, somewhat less coupling when participants are expected to integrate utilitarian assessments into their judgments, and the least coupling when participants are asked to make purely utilitarian assessments. These data provide critical information concerning the respective roles of the vmPFC and amygdala in moral judgment, and concerning the neural architecture of moral judgment more generally.

Some have argued that the vmPFC generates affective responses to behavioral options, for example, by binding stimuli with somatic markers (Damasio, 1996; Bechara and Damasio, 2005). Some claim that the vmPFC collects goal-relevant affective information (e.g., potential reinforcement) and integrates that information into subjective value signals that guide behavior (Rolls, 2005; Rangel and Hare, 2010; Grabenhorst and Rolls, 2011; Padoa-Schioppa and Cai, 2011). The latter view—whether one regards it as an alternative to, or an elaboration of, the former—makes a more specific prediction concerning how the vmPFC will engage with morally salient stimuli depending on the task. If the vmPFC merely triggers emotional states relevant to the task at hand, activity in this region should be similar if not greater when only considering the emotional aversiveness of two options (EA) rather than making integrative, "all things considered" moral judgments (IMJ) that include explicit rule-based reasoning. However, our data instead indicate that the vmPFC is preferentially engaged when emotional responses and explicit rule-based reasoning must be integrated to form an "all things considered" judgment. Likewise, we find that an overlapping region of vmPFC exhibits increased activity when judging "all things considered" moral acceptability compared with assessing well defined costs and benefits. This is consistent with a role for this region more generally in processes that are goal directed rather than purely Pavlovian on the one hand or purely rule bound on the other (Rangel et al., 2008; Balleine and O'Doherty, 2010).

These findings support the hypothesis that the vmPFC integrates disparate value signals into a more abstract, summary value representation, not only in the domain of personal economic choice, but in the domain of moral judgment concerning third parties. Previous research supports this view with respect to the integration of information concerning outcome magnitude and probability (Shenhav and Greene, 2010). The present results take this theory a step further, providing evidence consistent with the vmPFC's serving as a locus of integration for the two dominant modes of valuation in moral judgment—emotional evaluation supporting deontological judgment and the utilitarian

assessment of consequences. Because our design collects only one kind of behavioral response per trial, we cannot directly model the formation of integrative representations in the vmPFC (as in Shenhav and Greene, 2010). However, our finding that the vmPFC is most active in the integrative judgment condition (Fig. 3) is complemented by findings from our behavioral data (Fig. 2) and analyses of functional connectivity. These findings combine with the literature on economic decision-making (Wallis, 2007; Rangel and Hare, 2010; Schoenbaum and Esber, 2010; Grabenhorst and Rolls, 2011; Padoa-Schioppa and Cai, 2011) to paint a consistent picture of the vmPFC as an integrator of distinct, and sometimes conflicting, moral assessments.

Our findings are broadly consistent with the predictions of Blair (2007), who suggests that the vmPFC's role in moral judgment is integrative and modulated by present goals (rather than purely reactive) and that the amygdala, in contrast, plays a critical role in reacting to options that involve actively causing harm, potentially biasing moral judgments against such options. (For a more process-oriented account, see Schaich Borg et al., 2011). This prediction is also suggested by studies implicating the amygdala in moral judgment, though not in any specific kind of moral response (Greene et al., 2004; Glenn et al., 2009). Likewise, it is suggested by studies highlighting the amygdala's more general role in guiding behavior based on "bottom up," stimulus-based information that is emotionally relevant or salient to the decision-maker (Anderson and Phelps, 2001; Phelps and LeDoux, 2005; Seymour and Dolan, 2008). Consistent with this, recent accounts of dual-process moral cognition suggest that deontological judgments against harmful actions are driven by learned associations acquired through Pavlovian or "model-free" learning systems (Crockett, 2013; Cushman, 2013).

These findings have several implications for our understanding of how the human brain makes moral judgments, addressing three central questions in the field. First, how are internal conflicts between competing moral assessments resolved? Based on the previous literature, the answer is by no means obvious. On the one hand, it is now well established that patients with vmPFC damage make more utilitarian judgments than healthy controls (Ciaramelli et al., 2007; Koenigs et al., 2007; Moretto et al., 2010; Thomas et al., 2011). These results suggest a "partisan" role for the vmPFC, inherently favoring deontological judgment over utilitarian judgment when the two conflict. Conversely, as just noted, research on domain-general decision-making suggests that the vmPFC is a kind of level affective playing field on which incoming value signals compete for behavioral control. The present data are generally consistent with the latter view, but this leaves a puzzle: why does vmPFC damage lead specifically to more utilitarian judgment rather than to a general disruption in moral judgment with no particular behavioral bias? Our hypothesis is that explicit utilitarian reasoning can influence judgment independently of the vmPFC, which is consistent with the finding that vmPFC patients are perfectly capable of explicit reasoning (Saver and Damasio, 1991). We hypothesize, in contrast, that the negative signals produced by the amygdala in response to prototypically harmful actions (Greene et al., 2009) primarily exert their influence on voluntary, goal-directed behavior through the vmPFC. This hypothesis is consistent with our findings concerning coupling between amygdala and vmPFC under conditions requiring attention to emotional responses. Therefore, we hypothesize that vmPFC damage favors utilitarian judgment, not by damaging a "partisan" process, but by damaging an integration process that is necessary for one "partisan" process, but not its opponent, to influence judgment (cf. Schaich Borg et al., 2011).

Future research may provide a more conclusive test of this hypothesis.

Second, our analyses correlating amygdala response with behavioral ratings suggest that the amygdala plays an important role in generating emotional aversions to harmful actions and supports subsequent "all things considered" judgments against such actions in the face of countervailing utilitarian considerations. This provides the most direct neural evidence yet for models of moral judgment according to which emotional reactions play a crucial role in determining "all things considered" moral judgments (Haidt, 2001; Greene and Haidt, 2002).

Finally, our results help to answer a third key question in the cognitive neuroscience of moral judgment: to what extent do moral judgments depend on domain-specific versus domain-general processes (Greene and Haidt, 2002; Shenhav and Greene, 2010; Young and Dungan, 2012)? Some have argued that moral judgments depend critically on a domain-specific moral faculty (Hauser, 2006; Huebner et al., 2009; Mikhail, 2011). Although the present results do not conclusively rule out this hypothesis, they do count against it. Contrary to this strong modular view of moral cognition, the present results suggest that the amygdala and vmPFC, when confronted with a moral problem, perform functions consistent with those performed outside of the moral domain, with the amygdala playing a signaling role in response to salient features of stimuli and the vmPFC playing a more integrative role in response to task demands.

References

Anderson AK, Phelps EA (2001) Lesions of the human amygdala impair enhanced perception of emotionally salient events. Nature 411:305–309. CrossRef Medline

Balleine BW, O'Doherty JP (2010) Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology 35:48–69. CrossRef Medline

Bechara A, Damasio A (2005) The somatic marker hypothesis: a neural theory of economic decision. Games and Economic Behavior 52:336–372. CrossRef

Berthoz S, Grèzes J, Armony JL, Passingham RE, Dolan RJ (2006) Affective response to one's own moral violations. Neuroimage 31:945–950. CrossRef Medline

Blair RJ (2007) The amygdala and ventromedial prefrontal cortex in morality and psychopathy. Trends Cogn Sci 11:387–392. CrossRef Medline

Ciaramelli E, Muccioli M, Làdavas E, di Pellegrino G (2007) Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. Soc Cogn Affect Neurosci 2:84–92. 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

Crockett MJ (2013) Models of morality. Trends Cogn Sci 17:363–366. 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 FA, Greene JD (2011) Finding faults: how moral dilemmas illuminate cognitive structure. Soc Neurosci 7:269–279. CrossRef Medline

Cushman F, Murray D, Gordon-McKeon S, Wharton S, Greene JD (2012) Judgment before principle: engagement of the frontoparietal control network in condemning harms of omission. Soc Cogn Affect Neurosci 7:888–895. CrossRef Medline

Damasio AR (1996) The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philos Trans R Soc Lond, B, Biol Sci 351:1413–1420. CrossRef Medline

Deichmann R, Gottfried JA, Hutton C, Turner R (2003) Optimized EPI for fMRI studies of the orbitofrontal cortex. Neuroimage 19:430–441. CrossRef Medline

Foot P (1978) The problem of abortion and the doctrine of double effect. In: Virtues and vices. Oxford: Blackwell.

Gitelman DR, Penny WD, Ashburner J, Friston KJ (2003) Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. Neuroimage 19:200–207. CrossRef Medline

Glenn AL, Raine A, Schug RA (2009) The neural correlates of moral decision-making in psychopathy. Molecular Psychiatry 14:5–6. CrossRef Medline

Grabenhorst F, Rolls ET (2011) Value, pleasure and choice in the ventral prefrontal cortex. Trends Cogn Sci 15:56–67. CrossRef Medline

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

Greene JD, Cushman FA, Stewart LE, Lowenberg K, Nystrom LE, Cohen JD (2009) Pushing moral buttons: The interaction between personal force and intention in moral judgment. Cognition 111:364–371. 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

Harenski CL, Antonenko O, Shane MS, Kiehl KA (2010) A functional imaging investigation of moral deliberation and moral intuition. Neuroimage 49:2707–2716. CrossRef Medline

Hauser MD (2006) The liver and the moral organ. Soc Cogn Affect Neurosci 1:214–220. CrossRef Medline

Huebner B, Dwyer S, Hauser M (2009) The role of emotion in moral psychology. Trends Cogn Sci 13:1–6. CrossRef Medline

King JA, Blair RJR, Mitchell DG, Dolan RJ, Burgess N (2006) Doing the right thing: a common neural circuit for appropriate violent or compassionate behavior. Neuroimage 30:1069–1076. CrossRef Medline

Knutson B, Rick S, Wimmer GE, Prelec D, Loewenstein G (2007) Neural predictors of purchases. Neuron 53:147–156. CrossRef Medline

Koenigs M, Young L, Adolphs R, Tranel D, Cushman F, Hauser M, Damasio A (2007) Damage to the prefrontal cortex increases utilitarian moral judgements. Nature 446:908–911. CrossRef Medline

Kuhnen CM, Knutson B (2005) The neural basis of financial risk taking. Neuron 47:763–770. CrossRef Medline

Mendez MF, Anderson E, Shapira JS (2005) An investigation of moral judgement in frontotemporal dementia. Cogn Behav Neurol 18:193–197. CrossRef Medline

Mikhail J (2011) Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment. Cambridge: Cambridge UP.

Moretto G, Làdavas E, Mattioli F, di Pellegrino G (2010) A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. J Cogn Neurosci 22:1888–1899. CrossRef Medline

Padoa-Schioppa C, Cai X (2011) The orbitofrontal cortex and the computation of subjective value: consolidated concepts and new perspectives. Ann N Y Acad Sci 1239:130–137. CrossRef Medline

Paxton JM, Ungar L, Greene JD (2012) Reflection and reasoning in moral judgment. Cogn Sci 36:163–177. CrossRef Medline

Phelps EA, LeDoux JE (2005) Contributions of the amygdala to emotion processing: from animal models to human behavior. Neuron 48:175–187. CrossRef Medline

Rangel A, Hare T (2010) Neural computations associated with goal-directed choice. Curr Opin Neurobiol 20:262–270. CrossRef Medline

Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9:545–556. CrossRef Medline

Rissman J, Gazzaley A, D'Esposito M (2004) Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage 23:752–763. CrossRef Medline

Rolls ET (2005) Emotion explained. Oxford: OUP.

Saver JL, Damasio AR (1991) Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. Neuropsychologia 29:1241–1249. CrossRef Medline

Schaich Borg J, Hynes C, Van Horn J, Grafton S, Sinnott-Armstrong W (2006) Consequences, action, and intention as factors in moral judg-

ments: an FMRI investigation. J Cogn Neurosci 18:803–817. CrossRef Medline

Schaich Borg J, Sinnott-Armstrong W, Calhoun VD, Kiehl KA (2011) Neural basis of moral verdict and moral deliberation. Soc Neurosci 6:398–413. CrossRef Medline

Schoenbaum G, Esber GR (2010) How do you (estimate you will) like them apples? Integration as a defining trait of orbitofrontal function. Curr Opin Neurobiol 20:205–211. CrossRef Medline

Seymour B, Dolan R (2008) Emotion, decision making, and the amygdala. Neuron 58:662–671. 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

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

Thomas BC, Croft KE, Tranel D (2011) Harming kin to save strangers: further evidence for abnormally utilitarian moral judgments after ventrome-dial prefrontal damage. J Cogn Neurosci 23:2186–2196. CrossRef Medline

Thomson JJ (1986) Rights, restitution, and risk: essays in moral theory. Cambridge: Harvard.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15:273–289. CrossRef Medline

Wallis JD (2007) Orbitofrontal cortex and its contribution to decision-making. Annu Rev Neurosci 30:31–56. CrossRef Medline

Weiskopf N, Hutton C, Josephs O, Deichmann R (2006) Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. Neuroimage 33:493–504. CrossRef Medline

Young L, Dungan J (2012) Where in the brain is morality? Everywhere and maybe nowhere. Soc Neurosci 7:1–10. CrossRef Medline