Cerebral Cortex February 2009;19:276-283
doi:10.1093/cercor/bhn080
Advance Access publication May 22, 2008

The Neural Basis of Human Social Values: Evidence from Functional MRI

Roland Zahn[1,2], Jorge Moll[1,3], Mirella Paiva[1], Griselda Garrido[4], Frank Krueger[1], Edward D. Huey[1] and Jordan Grafman[1]

[1]National Institutes of Health, National Institutes of Neurological Disorders and Stroke, Cognitive Neuroscience Section, Bethesda, MD 20892-1440, USA, [2]The University of Manchester, School of Psychological Sciences, Neuroscience and Aphasia Research Unit, Manchester M13 9PL, United Kingdom, [3]Cognitive and Behavioral Neuroscience Unit, LABS-D'Or Hospital Network, 22280-080 Rio de Janeiro, Brazil, and [4]Instituto Israelita de Ensino e Pesquisa Albert Einstein, 05651-901 São Paulo, Brazil

Social values are composed of social concepts (e.g., "generosity") and context-dependent moral sentiments (e.g., "pride"). The neural basis of this intricate cognitive architecture has not been investigated thus far. Here, we used functional magnetic resonance imaging while subjects imagined their own actions toward another person (self-agency) which either conformed or were counter to a social value and were associated with pride or guilt, respectively. Imagined actions of another person toward the subjects (other-agency) in accordance with or counter to a value were associated with gratitude or indignation/anger. As hypothesized, superior anterior temporal lobe (aTL) activity increased with conceptual detail in all conditions. During self-agency, activity in the anterior ventromedial prefrontal cortex correlated with pride and guilt, whereas activity in the subgenual cingulate solely correlated with guilt. In contrast, indignation/anger activated lateral orbitofrontal-insular cortices. Pride and gratitude additionally evoked mesolimbic and basal forebrain activations. Our results demonstrate that social values emerge from coactivation of stable abstract social conceptual representations in the superior aTL and context-dependent moral sentiments encoded in fronto-mesolimbic regions. This neural architecture may provide the basis of our ability to communicate about the meaning of social values across cultural contexts without limiting our flexibility to adapt their emotional interpretation.

Keywords: Major Depression, Semantics, Moral Emotions, Anterior Temporal Lobe, Subgenual Cingulate Cortex

Introduction

We use social concepts (e.g., "honor," "generosity," "courage") to describe social, personal and moral values, also known as virtues. Social values are trans-situational goals which guide the evaluation of our own as well as other people's behavior (Schwartz and Bilsky 1987; Rohan 2000; Hitlin and Piliavin 2004) and consist of abstract conceptual knowledge linked to emotional states and social actions (Schwartz and Bilsky 1987). Due to their reliance on abstract conceptual knowledge, social values have a higher level of abstraction (Rohan 2000) than simple attitudes which are generally held to consist of emotional valence linked to a particular object, person or action (Rohan 2000; Cunningham and Zelazo 2007).

In a previous study, we identified a specialized superior anterior temporal lobe region (aTL, BA38/22) which represents abstract conceptual knowledge that enables us to comprehend social concepts (Zahn et al. 2007). How these neural representations of abstract social conceptual content are bound together with different contexts of social actions and emotions which dynamically shape our apprehension of social values is unknown. In this study, we investigated this issue using functional magnetic resonance imaging (fMRI).

One way of integrating the conceptual and emotional content of social values would be to directly link positive (reward) and negative (punishment) emotional valence to abstract conceptual representations. Following from this hypothesis, there should be valence-specific limbic brain activations which underpin feelings associated with social values independent of agency. Here, we will test an alternative model of integration of concepts and emotions that form social values (Moll, de Oliveira-Souza, Zahn, et al. 2007). According to this model, social values change their emotional quality in a flexible way adapted to the context of agency. Social values also have a stable core component which is their abstract conceptual meaning as expressed by social concepts (e.g., "honor", "courage") used to describe values across different personal and cultural contexts. This remarkable stability could be explained on the basis of abstract conceptual representations within the aTL which we hypothesized to be independent of contexts of emotions and actions (Zahn et al. 2007). This separation of stable context-independent representations in the aTL that can be flexibly embedded within different contexts of action implementation and emotional qualities as encoded in fronto-limbic circuits could account for our ability to link social values to a wide range of interpersonal and cultural settings.

The interdependency of context of actions and emotional evaluation has been a key component of the notion of values proposed by British philosophers during the 18th century. According to this stance, intuitive "moral sentiments" determine whether we perceive a behavior as constituting a virtue or vice and guide our approval or disapproval of that behavior (Hume 1777), a point of view which has gained recent support (Haidt 2001). Further, David Hume emphasizes the inextricable relation of actions as the objects of moral sentiments and notes that moral evaluation of such actions depends on whether these are caused internally or by external force. When we are the agent of an action conforming to our values, we may feel pride, whereas when another person is the agent, we may feel gratitude. On the negative side, when we act counter to our values, we may feel guilt and when another person acts in the same way toward us, we instead feel indignation or anger (Moll, de Oliveira-Souza, Zahn, et al. 2007).

Although we and others have referred to moral sentiments as "emotions," consistent evidence from functional imaging studies suggests that these complex subjective experiences arise from distributed activations in neocortical (anterior PFC and aTL) as well as phylogenetically older mesolimbic and orbitofrontal (OFC) regions (Moll et al. 2005). These findings

lead us to propose that moral sentiments emerge from the functional integration of activity in limbic regions encoding emotional states, PFC regions which represent event sequences and action outcomes (Wood and Grafman 2003), anterior temporal regions which represent abstract conceptual knowledge, and posterior temporal regions encoding sensory social features (Moll et al. 2005).

The distinction between different social concepts (e.g., "generosity," "honor," "politeness") lies in abstract conceptual descriptions of social behavior independent of the context of action and emotion (Zahn et al. 2007). Distinctions among different moral sentiments (e.g., "pride," "guilt," "gratitude," "indignation/anger") in contrast, are not defined by differences in abstract conceptual content, but by differences in contexts of agency and emotional states (Moll, de Oliveira-Souza, Zahn, et al. 2007). Social or moral values link abstract conceptual information to emotional flavors and contexts of action.

Here, we investigated the neural basis of social values by using the same abstract social concepts to evoke different qualities of moral sentiments through manipulating 2 important context variables of social value-related actions: self- versus other-agency and acting in accordance with, versus acting counter to, the social value described by an abstract concept.

Subjects underwent fMRI while they read sentences (e.g., "Tom [subject's own name] acts stingily [or generously] toward Sam [best friend's name]," "Sam acts stingily [or generously] toward Tom"). During the scan subjects judged the pleasantness of their own feelings associated with that behavior. To measure moral sentiments, subjects had to choose a label which best described their feelings related to the described social behaviors from their own perspective after the scan. The conditions were 1) self-agency in accordance with social values (positive, POS_S-AG), 2) other-agency in accordance with social values (positive, POS_O-AG), 3) self-agency counter to social values (negative, NEG_S-AG), 4) other-agency counter to social values (negative, NEG_O-AG). This design allowed us to carefully control the properties of stimuli used across the different conditions and to probe the abstract conceptual content as well as the emotional and action context of social values.

Because no previous study has compared positive and negative moral sentiments evoked by different agency-roles and abstract social concepts, our experimental hypotheses for categorical effects of different sentiments were based on drawing analogies to previous work on altruistic decisions during charity donation (Moll et al. 2006) and script-driven elicitation of moral sentiments (Moll et al. 2005; Moll, de Oliveira-Souza, Garrido, et al. 2007). We hypothesized that empathic prosocial sentiments (guilt) would activate the subgenual PFC and/or septum, regions recently implicated in social attachment and pair bonding in human and animal studies (Insel and Young 2001; Bartels and Zeki 2004; Depue and Morrone-Strupinsky 2005; Moll et al. 2006). For sentiments evoked during self-agency (pride, guilt), we expected stronger medial PFC activity necessary to predict outcomes of one's actions which determine the causal attribution of locus of agency to oneself necessary to evoke the feeling (Moll, de Oliveira-Souza, Zahn, et al. 2007). In addition, we expected predominantly lateral OFC-insular and dorsolateral PFC for other-critical sentiments (indignation/anger; Blair et al. 1999; Moll et al. 2006). Finally, for positive sentiments (pride and gratitude) related to value-guided social behavior, we predicted activation in regions within the mesolimbic reward pathway and its projections to the basal forebrain (ventral tegmental area [VTA], ventral striatum, hypothalamus, septum) grounded on the observed activations in this network during altruistic decisions and the hypothesized role of the basal forebrain in affiliative rewards.

Our results showed that the superior aTL is recruited during emotional judgments of social value-related behavior and that activity is indeed independent of valence and agency. Further, we demonstrated that different moral sentiments can be distinguished by differential activations within fronto-mesolimbic subregions. The prediction of predominantly medial PFC activity for moral sentiments evoked by self-agency (pride, guilt) and predominantly lateral PFC activity for other-critical sentiments (indignation/anger) was confirmed. Also the predictions of activity in different mesolimbic reward and basal forebrain regions (VTA, hypothalamus and septum) for gratitude and pride and the subgenual cingulate activation for guilt were substantiated. Neither valence nor agency alone accounted for categorical differences in activation within these regions corroborating our hypothesis that moral sentiments associated with social values cannot be explained solely on the basis of the main effects of these factors.

Materials and Methods

Subjects

Twenty-nine healthy subjects (15 men, age: mean = 27.9 ± 7.3 years, education: mean = 17.2 ± 1.5 years) took part in the fMRI experiment, none of whom had participated in our previous study (Zahn et al. 2007). Data from 5 additional subjects had to be excluded prior to the statistical analysis (N = 3, MR-scanner failure; N = 1, participant fell asleep; N = 1, ventromedial PFC signal loss). All were strongly right-handed and native English speakers, underwent a neurological examination and a clinical screening MRI during the previous 12 months, had normal or corrected-to-normal vision, no history of psychiatric or neurological disorders or psychopharmacological treatment, were not taking centrally active medications and had not consumed alcohol 24 h prior to scanning. Informed consent was obtained according to procedures approved by the NINDS Internal Review Board. Subjects were compensated for their participation according to the NINDS standards.

fMRI Paradigm

The 5 conditions during visually presented event-related fMRI were 1) POS_S-AG (N = 45), 2) POS_O-AG (N = 45), 3) NEG_S-AG (N = 45), 4) NEG_O-AG (N = 45), 5) fixation of visual pattern (FIX, null event, N = 90).

The social concepts were a subset of stimuli used in an independent previous study (Zahn et al. 2007) for which we had acquired normative data on the descriptiveness of social behavior. In the prestudy the degree of detail with which each concept described social behavior was assessed. We acquired additional normative data in N = 64 subjects (33 men, age: mean = 28.1 ± 7.7 years, education: mean = 17.3 ± 2.1 years, including the subjects of this study and the previous fMRI study (Zahn et al. 2007)) on familiarity from personal experience (1- to 7-point Likert Scale) and pleasantness/unpleasantness (–4 to +4 bipolar Likert scale) of each social behavior (N = 180) described by a statement. Subjects rated after the scan whether it was important for them 1) to act or 2) not to act in a way described by each social concept, or whether 3) it was not important.

Relevant psycholinguistic variables from the MRC Psycholinguistic database (Coltheart 1981): word familiarity, Kucera Francis word frequency, imageability and concreteness in addition to number of syllables were matched across conditions (see Supplementary Materials and Methods). The sentence structure and word number was identical for all stimuli.

Image Acquisition

Echoplanar T_2^*-weighted images were acquired (344 volumes per run) on a 3 Tesla General Electric scanner equipped with a standard head coil, high-order manual shimming to temporal and ventral frontal lobes, 3-mm slice thickness, 64 × 64 matrix, 37 slices, repetition time [TR] = 2.3 s, echo time [TE] = 20.5, field of view [FOV]: 220 × 220 mm, parallel to the anterior to posterior commissural line, whole brain coverage (not cerebellum). The first 5 volumes were discarded. The combination of high-field MRI, thinner slices (Bodurka et al. 2007) and high-order manual shimming optimized the signal in anterior temporal and ventral frontal lobes. All subjects had full coverage of the aTL and most of the ventral frontal cortex upon inspection of normalized echoplanar images (see Supplementary Fig. 7). In addition, high-resolution ($\approx$1 mm^3) T_1-weighted 3D magnetization-prepared rapid acquisition gradient echo structural images were collected (1-mm slice thickness, 128 slices, matrix: 224 × 224, TE = 2.964; FOV: 220 × 222 mm).

Image Analysis

Imaging data were analyzed using statistical parametric mapping (SPM5, http://www.fil.ion.ucl.ac.uk/spm/software/spm5) and a general linear model (Friston et al. 1995). For each condition, descriptiveness of social behavior of social concepts was modeled as a parametric predictor convolved with the hemodynamic response function (HRF). In addition, for each condition, the most frequently occurring moral sentiment was modeled as a categorical predictor (see Supplementary Fig. 5) for the respective condition for each subject. Finally, self-distinctiveness of social values (i.e., the difference of self-reference minus best-friend reference) was modeled as a parametric predictor of no interest for each condition and subject at the first level (for a schematic outline of the SPM5 analyses see Supplementary Fig. 9).

Familiarity from personal experience and pleasantness of social behaviors as well as self-reference and best-friend-reference of social concepts (rated as part of our normative study in N = 64 subjects for each of the 90 social concepts on 1- to 7-point Likert Visual Analog Scales: "How well does the word describe you [your best friend]?") used to describe these behaviors were so highly correlated that 94.8% of the total variance on all those variables could be explained by a single principle component (principle components analysis, SPSS14: http://www.spss.com) and each variable loaded onto this component with minimum correlations of 0.95. Therefore we decided not to include any of those highly correlated variables into our model and instead modeled the effect of valence categorically within our factorial model. In interpreting this and other studies, the potentially high correlation of self-reference, similar other-reference, valence and familiarity needs to be kept in mind (regarding the fallacies of including highly correlated variables as predictors in multiple regression models see Stevens 1996).

All partial effects of interest per condition were compared versus the low-level baseline (fixation): 1) condition-specific HRF, 2) descriptiveness of social behavior convolved with HRF, 3) condition-specific moral sentiment convolved with HRF. These contrasts were entered at the second-level using a random-effects factorial model with 2 categorical factors: 1) valence (positive/negative) and 2) agency (self/other) resulting in the 4 experimental conditions.

To reveal brain regions commonly activated across all conditions vs. Fixation, we performed a conjunction null analysis (Friston et al. 2005) on the sum of partial effects of interest (HRF, descriptiveness of social behavior, condition-specific moral sentiment) over all 4 conditions at an uncorrected P = 0.001, 5 voxels (corresponding to a false positive per voxel probability <0.01 according to Monte-Carlo simulations, Forman et al. 1995). Further we exclusively masked this conjunction analysis by F-tests for main effects and interactions of valence and agency at a lenient threshold (P = 0.05, uncorrected) to rigorously exclude voxels where there were significant differences across conditions. This masked conjunction analysis was used as a region of interest (ROI) mask (Maldjian et al. 2003) in subsequent analyses on the partial effect of interest: descriptiveness of social behavior (P = 0.001 uncorrected, 5 voxels) over all conditions, thereby revealing common regions where there was an increase of activity increasing with the respective predictor of interest. To test the spatial reliability of our finding for descriptiveness of social behavior, we created an inclusive mask using the effect of descriptiveness of social behavior from our first study (Zahn et al. 2007) at a lenient threshold (P = 0.05, 5 voxels within the same aTL ROI, created for our first study).

To examine the main effects and interactions of valence and agency (P = 0.001, 5 voxels), we masked the main effects exclusively by F-tests on the complementary main effect and the interaction at a lenient threshold (P = 0.05) to rigorously isolate voxels which solely respond to the main effect of interest.

In order to examine effects of condition-specific moral sentiment within in each condition, we carried out a separate analysis where condition-specific moral sentiment versus Fixation was analyzed with a subject-specific covariate of the Z-score for overall moral sentiment frequency of experience during the experiment per subject (see Supplementary Materials and Methods). This allowed us to test for effects consistent across subjects with the variance explained by individual variability on frequency of moral sentiment experience removed (covariance analysis) and separately to look at the effects of interindividual differences.

To increase the power of this more fine grained analysis we performed these analyses at an uncorrected P = 0.005, 4 voxels, which according to Monte-Carlo simulations (Forman et al. 1995) corresponds to a per voxel false positive probability of P = 0.01 to 0.02. Only regions which additionally survived a family-wise error (FWE)-corrected threshold of P = 0.05 over a priori ROIs or the whole brain and additional inclusive masking with 2 higher-level contrasts are reported (P = 0.05, 5 voxels uncorrected): 1) condition-specific moral sentiment versus same agency and opposite valence condition 2) condition-specific moral sentiment versus same valence and opposite agency condition. This inclusive masking was applied to focus the analysis on regions where there were categorical differences between the condition-specific moral sentiment in different conditions that could not be explained by the main effects of valence or agency.

We also looked for main effects of agency and valence on the condition-specific moral sentiment by inclusively masking simple contrasts for pride × POS_S-AG versus FIX and guilt × NEG_S-AG versus FIX at P = 0.005 by higher-level contrasts: pride × POS_S-AG versus gratitude × POS_O-AG and guilt × NEG_S-AG versus indignation/anger × NEG_O-AG at P = 0.05 to look for the effect of self-agency and the reverse contrasts for effects of other-agency. Simple contrasts of pride × POS_S-AG versus FIX and gratitude × POS_O-AG versus FIX at P = 0.005 were inclusively masked by contrasts at P = 0.05: pride × POS_S-AG versus guilt × NEG_S-AG and gratitude × POS_O-OAG versus indignation/anger × NEG_O-AG to reveal effects of positive valence and the reverse contrasts to look at negative valence effects.

To correct for multiple comparisons in all reported analyses, we created bilateral anatomical ROIs (see Supplementary Materials and Methods) for all a priori regions predicted to be relevant for social concepts, moral sentiments and agency (Moll et al. 2005; Moll, de Oliveira-Souza, Garrido, et al. 2007): aTL, posterior superior temporal sulcus/temporo-parietal junction [pSTS_TPJ], dorsolateral PFC, ventromedial PFC, lateral OFC, dorsomedial PFC, primary and supplementary motor cortex, insula, amygdala, basal ganglia, septum, hypothalamus, VTA). Only regions surviving FWE-corrected P = 0.05 over the bilateral predefined ROI volume were reported. Activations outside of a priori ROIs were reported when they survived a whole brain FWE-corrected threshold of P = 0.05. All reported coordinates are in Montreal Neurological Institute Standard Space. MRIcron (http://www.sph.sc.edu/comd/rorden/mricron/, Rorden and Brett 2000) was used to display saved statistical masks overlaid on a standard template. For confirmatory statistics performed on peak voxel parameter estimates we report 2-tailed significances.

Results

Behavioral Data

Mean response times were equal across the 4 experimental conditions (one-way analysis of variance [ANOVA], $F_{3,176} = 0.47$, P = 0.70, see Supplementary Fig. 4). A high proportion of social values expressed by positive concepts (e.g., "generosity") were rated as personally important goals to promote (mean = 72.3 ± 2.9%

standard deviation), whereas values expressed by negative social concepts (e.g., "stinginess") were rated as personally important goals to prevent (mean = 78.5 ± 3.0%), with no difference between the number of personally important items between the conditions (Wilcoxon test, $Z = -0.89$, asymptotic 2-tailed $P = 0.37$). Rated familiarity and pleasantness/unpleasantness (i.e., valence) were equal between self-agency and other-agency conditions (Supplementary Fig. 5).

As predicted, pride was the most frequent moral sentiment in the POS_S-AG condition, gratitude in the POS_O-AG, guilt in the NEG_S-AG and indignation/anger in the NEG_O-AG condition (Supplementary Fig. 5).

After the scan we assessed strategies used by the subjects during fMRI regarding retrieval of autobiographical episodes and visual imagery (adapted from Piefke et al. 2005) and modulating effects on activity related to individual differences in moral sentiments could be excluded (see Supplementary Materials and Methods).

Temporal Lobe Responses Common Across Conditions

There was a significant effect of descriptiveness of social behavior within the right superior aTL (Fig. 2*a*,*b*), which was independent of agency and valence and consistent across all conditions (Supplementary Fig. 6*a*). We also separately tested main effects and interactions of valence and agency on the partial effects of descriptiveness of social behavior and, as expected, no effects emerged within the aTL. The activation peak of the superior aTL region overlaps with the region identified as specific for social concepts vs. animal function concepts in our previous study (Zahn et al. 2007). To test the spatial reliability, we created a mask using the effect of descriptiveness of social behavior from (Zahn et al. 2007). Figure 2*b* shows the result of the partial effect of descriptiveness of social behavior common across conditions inclusively masked by the statistical mask for the descriptiveness of social behavior effect in (Zahn et al. 2007). The superior aTL cluster located at the border of anterior BA22 and superior BA38 was the only region surviving this analysis.

Categorical Effects of Moral Sentiments and Interindividual Differences

We first tested whether there were common regions for positive sentiments vs. negative sentiments or the reverse and whether there were common regions for self-agency vs. other-agency-related sentiments or the reverse (see Materials and Methods). No significant regions could be detected, demonstrating that differential activations cannot be explained by the simple effects of valence or agency.

Effects of Interindividual Differences

Individual differences in the percentage of trials where pride was experienced during fMRI only lead to a single significant region on the whole brain analysis: the septum (higher in people with a higher frequency of pride during pride-related trials: 0, 15, 6; $Z = 3.18$, set-level corrected $P = 0.009$; peak voxel correlation: $R = 0.46$, $P = 0.01$, Supplementary Table 1, Fig. 1*c*, Fig. 3*a*,*b*).

Higher frequency of guilt was correlated with activity in anterior ventromedial PFC (BA10; –21, 51, –3; $Z = 3.67$) and the subgenual cingulate (BA32; –15, 36, –6; $Z = 5.48$, Supplementary Table 1, Fig. 1*f*, Fig. 3*a*,*c*). Individual difference effects within the subgenual cingulate cortex (BA32) for guilt ($R = 0.66$, $P < 0.0001$) were significantly higher compared with the other moral sentiments and there were no significant correlations with the other moral sentiments in this region (Fig. 3).

Higher individual frequency of gratitude was solely correlated with the hypothalamus in the whole brain analysis (3, –3, –3; $Z = 4.04$, Fig. 1*d*, Supplementary Table 1). No individual difference effects in the whole brain analysis could be detected for indignation/anger.

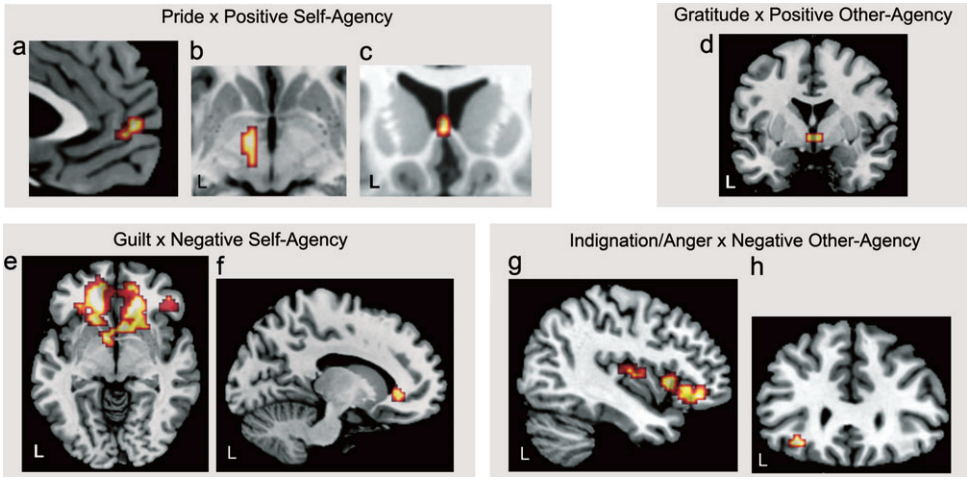

Figure 1. The partial effects for each condition-specific moral sentiment compared with Fixation are displayed (see also Materials and Methods and Supplementary Table 1). These are inclusively masked by 2 contrasts: 1) comparing the condition-specific moral sentiment (e.g., pride during POS_S-AG) versus the condition with opposite valence and same agency role (e.g., guilt during NEG_S-AG). 2) comparing versus the condition with opposite agency role and same valence (e.g., gratitude during POS_O-AG). This applies to all depictions of whole brain analyses. Additional a priori ROI analyses were carried out on the simple contrasts versus fixation without applying inclusive masking and all reported regions survived FWE-corrected $P = 0.05$ over a priori ROIs. All images are displayed at an uncorrected $P = 0.005$, 4 voxels. Consistent group effects for pride: (*a*) whole brain, (*b*) VTA ROI analysis, (*c*) effect of individual differences for pride, whole brain, (*d*) individual difference effect for gratitude, whole brain, (*e*) individual difference effect for guilt, overlay of ventromedial PFC, lateral OFC, basal ganglia, and septum ROIs. (*f*) Individual difference effect for guilt, ventromedial PFC ROI, masked by contrasts versus pride and indignation/anger. (*g*) Consistent group effect for indignation/anger, overlay of lateral OFC and insula ROIs. (*h*) Consistent group effect for indignation/anger, whole brain (see Supplementary Materials and Methods on definition of ROIs).

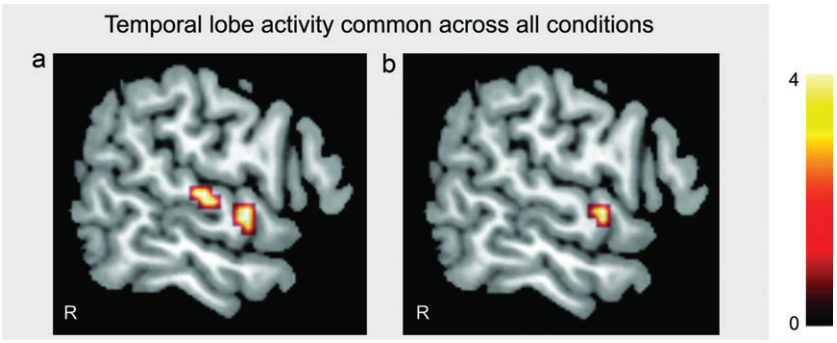

Figure 2. (*a*) Partial effect of descriptiveness of social behavior common across all conditions: right superior aTL (BA22, $x = 54$, $y = 0$, $z = -3$; $T = 4.53$; FWE-corrected P over a priori ROI = 0.008) and right superior mid-posterior temporal gyrus (BA22, $x = 57$, $y = -18$, $z = 6$; $T = 4.6$; FWE-corrected P over a priori ROI = 0.008). (*b*) The same analysis (as in *a*) inclusively masked by the descriptiveness of social behavior effect in Zahn et al. (2007) within the aTL ROI in right superior aTL (BA22, $x = 60$, $y = -3$, $z = -3$; $T = 4.18$; FWE-corrected P over a priori mask = 0.005). Activations are displayed at uncorrected $P = 0.001$, 5 voxels.

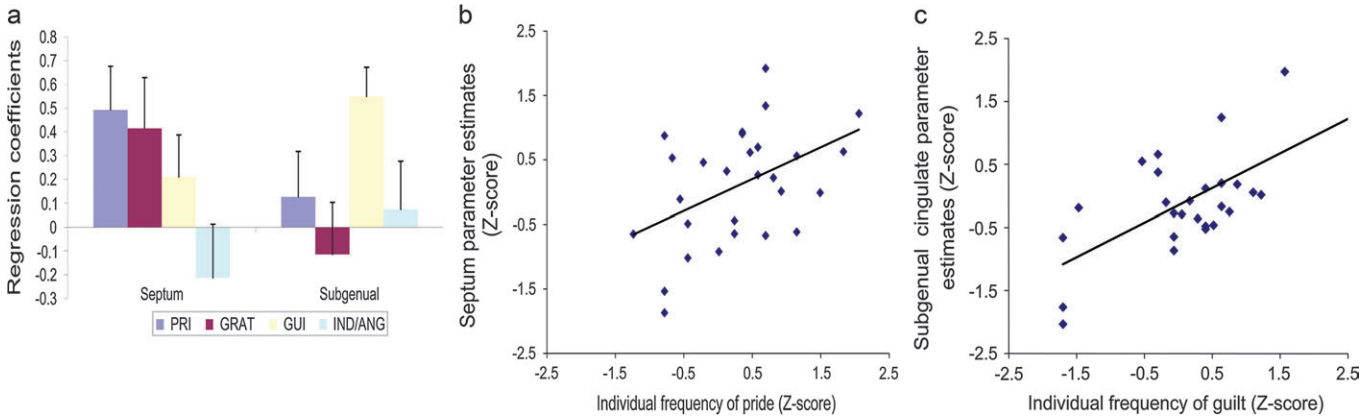

Figure 3. (*a*) Regression coefficients for subject-specific moral sentiment frequency covariate effects and standard errors in septum and subgenual cingulate (BA32) peak voxels. Scatter plot for Z-transformed fMRI effects for (*b*) pride in the septum and (*c*) guilt in the subgenual cingulate. There was a significant interaction of condition and effect of subject-specific moral sentiment covariates on Z-transformed fMRI effects within the septum (peak coordinate from Supplementary Table 1, univariate ANOVA, SPSS14, outliers Z-score outside ± 2.5 excluded): $F_{107,7} = 3.18$, $P = 0.03$. There was also a significant main effect of the moral sentiment covariate on the septal signal strength $F_{114,1} = 4.61$, $P = 0.03$. Unadjusted correlations for septal activity with pride for POS_S-AG: $R(29) = 0.46$, $P = 0.01$ (trend for gratitude: $R(29) = 0.35$, $P = 0.06$, negative trend for indignation/anger: $R(28) = -0.34$, $P = 0.08$, no significant correlations with guilt: $R(29) = 0.22$, $P = 0.25$). There was a significant interaction of condition with the moral sentiment covariate effect on the signal within the subgenual PFC: $F_{112,3} = 2.63$, $P = 0.05$. Subgenual PFC × NEG_S-AG_guilt: $R(28) = 0.66$, $P < 0.0001$ (no significant correlations with other moral sentiments: $P > 0.30$). There was a significant correlation of gratitude with hypothalamic activity ($R = 0.43$, $P = 0.02$). However, when adjusting the correlation for the effects of the other conditions, the overall ANOVA shows no significant effect of condition for the strength of correlation between moral sentiment covariates and hypothalamic activation and also no interaction of condition and moral sentiment covariates (at $P = 0.05$). Thus the effects in the hypothalamus were not robust enough to survive adjustment on the secondary data analysis.

Consistent Group Effects with Covariance due to Interindividual Differences Removed

Signal increases for pride within anterior ventromedial PFC (BA10; –9, 54, –3; $Z = 4.54$, Fig. 1*a*, Supplementary Table 1), the VTA (–9, –9, –6; $Z = 4.26$; Fig. 1*b*, Supplementary Table 1, reaching into the posterior hypothalamus) and the parahippocampal gyrus (BA30; –9, –48, 3; $Z = 5.82$; Supplementary Table 1) were consistent across subjects.

There were no interindividually consistent group effects which were specific for gratitude compared with pride and indignation/anger.

Indignation/anger evoked strong consistent group increases in activity within left OFC (BA47; –30, 30, –12; $Z = 4.81$, Fig. 1*g*,*b*, Supplementary Table 1), anterior insula (–36, 15, 3; $Z = 3.94$; Fig. 1*g*, Supplementary Table 1) and left dorsolateral PFC (BA9; –45, 6, 30; $Z = 4.16$, Supplementary Table 1). There were no regions in which guilt evoked stronger effects consistent across the group than indignation/anger and pride.

Discussion

In summary, we confirmed the hypothesis that social values draw upon stable representations of conceptual detail within the superior aTL and context-dependent representations of distinct moral sentiments within fronto-mesolimbic regions.

Temporal Lobe Responses Common Across Conditions

As predicted, the same right superior aTL region previously shown to represent abstract conceptual social knowledge (Zahn et al. 2007) was also recruited during emotional judgments of social values and this activation was independent of valence and agency. Superior aTL activity was not only associated with the richness of detail with which concepts describe social behavior but also with the subjective experience of moral sentiments during evaluations of social value-related actions described by these concepts (Supplementary Fig. 6*b*). This supports the notion that the experience of moral sentiments partly depends on abstract conceptual

representations of social behaviors within the aTL (Zahn et al. 2007). Further we confirmed our second hypothesis that categorical differences between moral sentiments evoked by evaluation of social value-related behavior are based on distinct patterns of fronto-mesolimbic brain activity which cannot be explained by valence effects alone.

Several regions showed significant activations common across all conditions and were more active for more descriptive concepts in addition to the superior aTL (Supplementary Table 2). Activity in none of those regions was detected in association with abstract conceptual knowledge of social behaviors tested during our previous fMRI study (Zahn et al. 2007). Our assertion that the superior aTL represents abstract conceptual knowledge of social values relies thus on the analysis of both studies in conjunction (Fig. 3*b*).

Categorical Effects of Moral Sentiments and Interindividual Differences

Categorical effects of different moral sentiments as part of the context-dependent experience of social values occurred as significant interindividual difference effects and as effects consistent across the group. Interindividual differences in the frequency of experiencing guilt strongly predicted activity in the subgenual cingulate cortex. Previous neuroimaging studies investigating script-driven elicitation of guilt without modeling individual differences have failed to show activity in the subgenual cingulate cortex (Shin et al. 2000; Takahashi et al. 2004; Moll, de Oliveira-Souza, Garrido, et al. 2007). However, this region has been previously associated with altruistic behavior (Moll et al. 2006) and was therefore predicted to be activated during the experience of prosocial sentiments (Moll, de Oliveira-Souza, Garrido, et al. 2007). Further, resting state activity in this region is abnormal in patients with major depression (Drevets 2000; Mayberg et al. 2005), a disorder associated with overgeneralized sentiments of interpersonal guilt (O'Connor et al. 2002).

Individual differences in pride when acting in accordance with one's own values toward one's best friend (a prosocial form of pride) were associated with activity in the septum, a region recently implicated in pair bonding, affiliative reward and learning (Insel and Young 2001; Depue and Morrone-Strupinsky 2005; Moll et al. 2006). The cingulate gyrus, lateral septal nuclei, medial preoptic area, mediobasal hypothalamus and VTA form a neural system implicated in pair bonding and affiliative rewards across a broad range of species (Insel and Young 2001; Depue and Morrone-Strupinsky 2005). This system is modulated in part by oxytocin, recently shown to increase trust in human interactions (Kosfeld et al. 2005). Additional activity within the VTA during experience of pride concurs with the role of the VTA for basic (Tobler et al. 2005) and affiliative rewards.

Thus commensurate with our hypotheses, positive value-related moral sentiments activated subregions of the meso-limbic reward system and the basal forebrain. Activity for prosocial sentiments compared with indignation/anger was higher in basal forebrain regions (septum) and paralimbic cortex (subgenual cingulate) previously related to affiliative bonding (Insel and Young 2001; Bartels and Zeki 2004; Moll et al. 2006).

In keeping with our expectation, activity in anterior ventromedial PFC (BA10) increased for sentiments evoked by self-agency during anticipated value-related behaviors (pride and guilt). Anterior medial PFC (BA10) activation was consistently found for moral sentiments (Moll et al. 2005) and was previously demonstrated for guilt (Moll, de Oliveira-Souza, Garrido, et al. 2007). Patients with damage to this region together with lesions of more posterior ventromedial PFC show reduced guilt and compassion (Koenigs et al. 2007), inappropriate pride (Beer et al. 2006), a lack of empathic concern, increased irritability, impoverishment of feelings, difficulties making real-world decisions and adjusting their social behavior to the sequential context of actions which may manifest as social inappropriateness or reductions of motivated behavior (Eslinger and Damasio 1985; Anderson et al. 2006; Rankin et al. 2006; Eslinger et al. 2007). Furthermore, anterior ventromedial PFC was found to underlie sequence judgments on component events of complex daily life event sequences (e.g., "going to the restaurant"; "attending a funeral") during fMRI (Krueger et al. 2007) and patients with lesions encompassing this region have impairments in knowledge of sequences of actions (Zalla et al. 2003). Importantly, fMRI activations for event sequences were independent of emotional valence (Krueger et al. 2007). Thus activity for pride and guilt in this region is not explained by emotional states represented within anterior medial PFC (BA10). This finding is, however, compatible with the view that moral sentiments partly depend on representations of sequential outcomes of one's own and other people's actions (Moll et al. 2005).

In summary, both self-agency conditions elicited activations within ventromedial PFC regions, however with different anatomical distributions. Importantly, these activations cannot be explained by self-reference (i.e., the degree to which subjects thought that social concepts were characteristic of themselves, see Materials and Methods) a variable which was controlled for that has been consistently linked to medial PFC activity (Northoff et al. 2006). Rated self-reference of social concepts in our study was so highly correlated with reference to the best friend, positive valence (i.e., pleasantness) and familiarity that these variables were statistically inseparable (see Materials and Methods). Consequently, if anterior vmPFC activity had been due to self-reference, one would have expected that in the conditions using positive social concepts leading to pride and gratitude one should have seen higher medial PFC activity than in the negative conditions (being lower in self-reference). On the contrary, however, both positive and negative self-agency conditions leading to pride and guilt induced higher anterior vmPFC activity compared with the other conditions. This means that main effects of self-reference of social concepts were not enhancing activity in the anterior vmPFC but that there was an interaction with the context of agency role (self vs. other) which determined whether this region was activated.

Prior investigations of the immediate sense of self-agency during motor actions revealed the importance of motor, premotor and dorsomedial PFC regions (Frith et al. 2000; David et al. 2006) and the specific role of the temporo-parietal junction in distinguishing self and other during self-agency (Sirigu et al. 1999; Decety and Sommerville 2003; Decety and Grezes 2006). The lack of main effects of self- or other-agency on these regions in our study points to partially dissociable neural systems which underlie the immediate sense of self-agency during motor actions probed in previous studies and the representations of self-agency during complex value-related social behavior which were addressed here.

Indignation/anger was associated with prominently left lateral OFC and dorsolateral PFC activity as well as anterior insula activations. Activations of the anterior insula have been repeatedly demonstrated with aversive stimuli (Seymour et al. 2007). Lateral PFC activations accord with the notion that anger irrespective of valence and punishment associations is represented in the lateral OFC (Blair et al. 1999) and that lateral PFC regions are more important when a change in strategy or response is required under unexpected circumstances (Elliott et al. 2000; Wood and Grafman 2003; Kringelbach and Rolls 2004). Guilt also activated lateral OFC (BA47) regions but not as strongly as indignation/anger. These different effects for guilt and indignation/anger cannot be explained by differences in valence, because the percentage of subjects experiencing indignation/anger during the other-agency conditions was equally strongly negatively correlated with pleasantness (R = −0.85, P < 0.0001) as it was with guilt (R = −0.86, P < 0.0001) during the self-agency conditions.

Conclusions

Taken together, our findings show that the same superior aTL regions which represent abstract social concepts are recruited during emotional judgment of social values and are stable across different contexts of moral sentiments. Further, we showed categorical differences in fronto-mesolimbic regions for moral sentiments evoked by social value-related actions.

Our results are most parsimoniously explained by the assumption that social values emerge from coactivation of abstract conceptual representations within the superior aTL, emotional states represented in mesolimbic and basal forebrain regions (hypothalamus, septum, VTA, anterior insula) and emotion–action associations in OFC as well as sequential action outcomes in anterior medial PFC regions (Moll et al. 2005). Irrespective of the postulated function of subdivisions within fronto-mesolimbic circuits, our results demonstrate that differences in patterns of fronto-mesolimbic activity are associated with different subjective qualities of moral sentiments evoked by the same abstract conceptual content of social values in different contexts of action.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/

Funding

National Institute of Neurological Disorders and Stroke intramural funding to J.G.; German Academy of Natural Scientists Leopoldina Fellowship funded by the Federal Ministry of Education and Research (BMBF-LPD 9901/8-122) to R.Z; Brazilian Fundação de Amparo à Pesquisa do Estado de São Paulo grant (03/11794-6) supported G.G; LABS-D'Or Hospital Network, Rio de Janeiro, Brazil, supported J.M.

Notes

We thank Eric Wassermann for performing neurological exams, Kris Knutson, and several SPM experts from the discussion list for imaging analysis advice. *Conflict of Interest*: None declared.

Address correspondence to Jordan Grafman, PhD, NIH/NINDS, Cognitive Neuroscience Section, 10 Center Drive, Room 7D43, Bethesda, MD 20892-1440, USA. Email: grafmanj@ninds.nih.gov.

References

Anderson SW, Barrash J, Bechara A, Tranel D. 2006. Impairments of emotion and real-world complex behavior following childhood- or adult-onset damage to ventromedial prefrontal cortex. J Int Neuropsychol Soc. 12:224–235.

Bartels A, Zeki S. 2004. The neural correlates of maternal and romantic love. Neuroimage. 21:1155–1166.

Beer JS, John OP, Scabini D, Knight RT. 2006. Orbitofrontal cortex and social behavior: integrating self-monitoring and emotion-cognition interactions. J Cogn Neurosci. 18:871–879.

Blair RJR, Morris JS, Frith CD, Perrett DI, Dolan RJ. 1999. Dissociable neural responses to facial expressions of sadness and anger. Brain. 122:883–893.

Bodurka J, Ye F, Petridou N, Murphy K, Bandettini PA. 2007. Mapping the MRI voxel volume in which thermal noise matches physiological noise-implications for fMRI. Neuroimage. 34:542–549.

Coltheart M. 1981. The MRC psycholinguistic database. The quarterly journal of experimental psychology A.. Hum Exp Psychol. 33:497–505.

Cunningham WA, Zelazo PD. 2007. Attitudes and evaluations: a social cognitive neuroscience perspective. Trends Cogn Sci. 11:97–104.

David N, Bewernick BH, Cohen MX, Newen A, Lux S, Fink GR, Shah NJ, Vogeley K. 2006. Neural representations of self versus other: visual-spatial perspective taking and agency in a virtual ball-tossing game. J Cogn Neurosci. 18:898–910.

Decety J, Grezes J. 2006. The power of simulation: imagining one's own and other's behavior. Brain Res. 1079:4–14.

Decety J, Sommerville JA. 2003. Shared representations between self and other: a social cognitive neuroscience view. Trends Cogn Sci. 7:527–533.

Depue RA, Morrone-Strupinsky JV. 2005. A neurobehavioral model of affiliative bonding: implications for conceptualizing a human trait of affiliation. Behav Brain Sci. 28:313–350.

Drevets WC. 2000. Functional anatomical abnormalities in limbic and prefrontal cortical structures in major depression. Prog Brain Res. 126:413–431.

Elliott R, Dolan RJ, Frith CD. 2000. Dissociable functions in the medial and lateral orbitofrontal cortex: evidence from human neuroimaging studies. Cereb Cortex. 10:308–317.

Eslinger PJ, Damasio AR. 1985. Severe disturbance of higher cognition after bilateral frontal-lobe ablation—patient EVR. Neurology. 35:1731–1741.

Eslinger PJ, Moore P, Troiani V, Antani S, Cross K, Kwok S, Grossman M. 2007. Oops! Resolving social dilemmas in frontotemporal dementia. J Neurol Neurosurg Psychiatry. 78:457–460.

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC. 1995. Improved assessment of significant activation in functional magnetic-resonance-imaging (fMRI)—use of a cluster-size threshold. Magn Reson Med. 33:636–647.

Friston KJ, Frith CD, Turner R, Frackowiak RS. 1995. Characterizing evoked hemodynamics with fMRI. Neuroimage. 2:157–165.

Friston KJ, Penny WD, Glaser DE. 2005. Conjunction revisited. Neuroimage. 25:661–667.

Frith CD, Blakemore SJ, Wolpert DM. 2000. Abnormalities in the awareness and control of action. Philos Trans R Soc Lond B Biol Sci. 355:1771–1788.

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

Hitlin S, Piliavin JA. 2004. Values: reviving a dormant concept. Annu Rev Sociol. 30:359–393.

Hume D. 1777. An enquiry into the principles of morals. London: T. Cadell.

Insel TR, Young LJ. 2001. The neurobiology of attachment. Nat Rev Neurosci. 2:129–136.

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.

Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E. 2005. Oxytocin increases trust in humans. Nature. 435:673–676.

Kringelbach ML, Rolls ET. 2004. The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. Prog Neurobiol. 72:341–372.

Krueger F, Moll J, Zahn R, Heinecke A, Grafman J. 2007. Event frequency modulates the processing of daily life activities in human medial prefrontal cortex. Cereb Cortex. 17:2346–2353.
Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH. 2003. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage. 19:1233–1239.
Mayberg HS, Lozano AM, Voon V, McNeely HE, Seminowicz D, Hamani C, Schwalb JM, Kennedy SH. 2005. Deep brain stimulation for treatment-resistant depression. Neuron. 45:651–660.
Moll J, de Oliveira-Souza R, Garrido GG, Bramati IE, Caparelli-Daquer EMA, Paiva MLMF, Zahn R, Grafman J. 2007. The self as a moral agent: linking the neural bases of social agency and moral sensitivity. Social Neurosci. 2:336–352.
Moll J, de Oliveira-Souza R, Zahn R, Grafman J. 2007. The cognitive neuroscience of moral emotions. In: Sinnott-Armstrong W, editor. Morals in the brain: emotion, disease and development. Cambridge (MA): Massachusetts Institute of Technology Press. p. 1–17.
Moll J, Krueger F, Zahn R, Pardini M, de Oliveira-Souza R, Grafman J. 2006. Human fronto-mesolimbic networks guide decisions about charitable donation. Proc Natl Acad Sci USA. 103: 15623–15628.
Moll J, Zahn R, de Oliveira-Souza R, Krueger F, Grafman J. 2005. The neural basis of human moral cognition. Nat Rev Neurosci. 6:799–809.
Northoff G, Heinzel A, de Greck M, Bermpohl F, Dobrowolny H, Panksepp J. 2006. Self-referential processing in our brain–a meta-analysis of imaging studies on the self. Neuroimage. 31:440–457.
O'Connor LE, Berry JW, Weiss J, Gilbert P. 2002. Guilt, fear, submission, and empathy in depression. J Affect Disord. 71:19–27.
Piefke M, Weiss PH, Markowitsch HJ, Fink GR. 2005. Gender differences in the functional neuroanatomy of emotional episodic autobiographical memory. Hum Brain Mapp. 24:313–324.
Rankin KP, Gorno-Tempini ML, Allison SC, Stanley CM, Glenn S, Weiner MW, Miller BL. 2006. Structural anatomy of empathy in neurodegenerative disease. Brain. 129:2945–2956.
Rohan MJ. 2000. A rose by any name? The values construct. Person Soc Psychol Rev. 4:255–277.
Rorden C, Brett M. 2000. Stereotaxic display of brain lesions. Behav Neurol. 12:191–200.
Schwartz SH, Bilsky W. 1987. Toward a universal psychological structure of human-values. J Person Social Psychol. 53:550–562.
Seymour B, Singer T, Dolan R. 2007. The neurobiology of punishment. Nat Rev Neurosci. 8:300–311.
Shin LM, Dougherty DD, Orr SP, Pitman RK, Lasko M, Macklin ML, Alpert NM, Fischman AJ, Rauch SL. 2000. Activation of anterior paralimbic structures during guilt-related script-driven imagery. Biol Psychiatry. 48:43–50.
Sirigu A, Daprati E, Pradat-Diehl P, Franck N, Jeannerod M. 1999. Perception of self-generated movement following left parietal lesion. Brain. 122(Pt 10):1867–1874.
Stevens J. 1996. Applied multivariate statistics for the social sciences. Mahwah (NJ): Lawrence Erlbaum Associates.
Takahashi H, Yahata N, Koeda M, Matsuda T, Asai K, Okubo Y. 2004. Brain activation associated with evaluative processes of guilt and embarrassment: an fMRI study. Neuroimage. 23:967–974.
Tobler PN, Fiorillo CD, Schultz W. 2005. Adaptive coding of reward value by dopamine neurons. Science. 307:1642–1645.
Wood JN, Grafman J. 2003. Human prefrontal cortex: processing and representational perspectives. Nat Rev Neurosci. 4:139–147.
Zahn R, Moll J, Garrido G, Krueger F, Huey ED, Grafman J. 2007. Social concepts are represented in the superior anterior temporal cortex. Proc Natl Acad Sci USA. 104:6430–6435.
Zalla T, Pradat-Diehl P, Sirigu A. 2003. Perception of action boundaries in patients with frontal lobe damage. Neuropsychologia. 41:1619–1627.