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], 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, UK

[3] ***Cognitive and Behavioral Neuroscience Unit, LABS-D'Or Hospital Network, 22280-080 - Rio de Janeiro, RJ, Brazil***

[4] Instituto Israelita de Ensino e Pesquisa Albert Einstein, 05651-901 – São ***Paulo, SP, Brazil***

This article has been accepted for publication in Cerebral Cortex ©: 2008

**** To whom correspondence should be addressed:***

NIH/NINDS, Cognitive Neuroscience Section, 10 Center Drive, Room 7D43, Bethesda, MD 20892-1440, USA, grafmanj@ninds.nih.gov

Running head: The Neural Basis of Social Values

Abstract

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 fMRI while subjects imagined their own actions towards 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 towards 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 activity increased with conceptual detail in all conditions. During self-agency, activity in the anterior ventromedial prefrontal cortex correlated with pride and guilt while 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 co-activation of stable abstract social conceptual representations in the superior anterior temporal lobe 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.

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 (Hitlin and Piliavin 2004; Rohan 2000; Schwartz and Bilsky 1987) 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 (Cunningham and Zelazo 2007; Rohan 2000).

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 MRI.

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 et al. 2007b). 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 towards us, we instead feel indignation or anger (Moll et al. 2007b).

While 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 et al. 2007b). 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 two 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] towards Sam [best friend’s name]’, ‘Sam acts stingily [or generously] towards 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.

Since 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. 2007a; Moll et al. 2005). 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 (Bartels and Zeki 2004; Depue and Morrone-Strupinsky 2005; Insel and Young 2001; 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 et al. 2007b). 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.

Results

Behavioral data

Mean response times (RT) were equal across the four experimental conditions (one-way analysis of variance (ANOVA), $F[3,176]=.47$, $P=.70$, 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, $Z=-.89$, asymptotic 2-tailed $P=.37$). 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 1a,b), which was independent of agency and valence and consistent across all conditions (Supplementary Fig 6a). 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). Fig 1b 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=.009; peak voxel correlation: R=.46, P=.01, Supplementary Tab 1, Fig 1c, Fig 3a,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 Tab 1, Fig 1f, Fig 3a,c). Individual difference effects within the subgenual cingulate cortex (BA32) for guilt (R=.66, P<.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 1d, Supplementary Tab 1). No individual difference effects in the whole brain analysis could be detected for indignation/anger.

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 1a, Supplementary Tab 1), the VTA (-9, -9, -6; Z=4.26; Fig 1b; Supplementary Tab 1, reaching into the posterior hypothalamus) and the parahippocampal gyrus (BA30; -9, -48, 3; Z=5.82; Supplementary Tab 1) were consistent across subjects.

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

Indignation/anger evoked strong consistent group increases in activity within left OFC (BA47; -30, 30, -12; Z=4.81, Fig 1g&h, Supplementary Tab 1), anterior insula (-36, 15, 3; Z=3.94; Fig 1g, Supplementary Tab 1) and left dorsolateral PFC (BA9; -45, 6, 30; Z=4.16, Supplementary Tab 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 6b). 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 3b).

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 (Moll et al. 2007a; Shin et al. 2000; Takahashi et al. 2004). 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 et al. 2007a). 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 towards 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 (Depue and Morrone-Strupinsky 2005; Insel and Young 2001; 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 (Depue and Morrone-Strupinsky 2005; Insel and Young 2001). 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 mesolimbic reward system and the basal forebrain. Activity for prosocial

sentiments compared to indignation/anger was higher in basal forebrain regions (septum) and paralimbic cortex (subgenual cingulate) previously related to affiliative bonding (Bartels and Zeki 2004; Insel and Young 2001; 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 et al. 2007a). 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 (Anderson et al. 2006; Eslinger and Damasio 1985; Eslinger et al. 2007; Rankin et al. 2006). 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 (David et al. 2006; Frith et al. 2000) and the specific role of the temporo-parietal junction in distinguishing self and other during self-agency (Decety and Grezes 2006; Decety and Sommerville 2003; Sirigu et al. 1999). 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; Kringelbach and Rolls 2004; Wood and Grafman 2003). 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 ($R=-.85$, $P<.0001$) as it was with guilt ($R=-.86$, $P<.0001$) 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 co-activation 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.

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 pre-study 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

Echo-planar T2*-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 x 64 matrix, 37 slices, TR = 2.3 s, TE = 20.5, FOV: 220 x 220 mm, parallel to the anterior to posterior commissural line, whole brain coverage (not cerebellum). The first five 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 (≈1 mm^3) T1-weighted 3D Magnetization-Prepared Rapid Acquisition Gradient Echo structural images were collected (1 mm slice thickness, 128 slices, matrix: 224 x 224, TE = 2.964; FOV: 220 x 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 .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: a) valence (positive / negative) and b) 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 four conditions at an uncorrected P=.001, 5 voxels (corresponding to a false positive per voxel probability <.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=.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=.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=.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=.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=.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 vs. 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=.005, 4 voxels, which according to Monte-Carlo simulations (Forman et al. 1995) corresponds to a per voxel false positive probability of p=.01 to .02. Only regions which additionally survived a Family-Wise-Error (FWE)-corrected threshold of P=.05 over apriori ROIs or the whole brain and additional inclusive masking with two higher-level contrasts are reported (P=.05, 5 voxels uncorrected): 1) CONDITION-SPECIFIC MORAL SENTIMENT vs. same agency & opposite valence condition 2) CONDITION-SPECIFIC MORAL SENTIMENT vs. same valence & 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 x POS_S-AG vs. FIX and Guilt x NEG_S-AG vs. FIX at P=.005 by higher-level contrasts: Pride x POS_S-AG vs. Gratitude x POS_O-AG and Guilt x NEG_S-AG vs. Indignation/Anger x NEG_O-AG at P=.05 to look for the effect of self-agency and the reverse contrasts for effects of other-agency. Simple contrasts of Pride x POS_S-AG vs. FIX and Gratitude x POS_O-AG vs. FIX at P=.005 were inclusively masked by contrasts at P=.05: Pride x POS_S-AG vs. Guilt x NEG_O-AG and Gratitude x POS_O-OAG vs. Indignation/Anger x 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 apriori regions predicted to be relevant for social concepts, moral sentiments and agency (Moll et al. 2007a; Moll et al. 2005): 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=.05 over the bilateral predefined ROI volume were reported. Activations outside of apriori regions of interest were reported when they survived a whole brain FWE-corrected threshold of P=.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.

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 Neuropsych 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. Journal of Cognitive Neuroscience. 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, Human experimental psychology. 33:497-505.

Cunningham WA, Zelazo PD. 2007. Attitudes and evaluations: A social cognitive neuroscience perspective. Trends in Cognitive Sciences. 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, Sommerville JA. 2003. Shared representations between self and other: A social cognitive neuroscience view. Trends Cogn Sci. 7:527-533.

Decety J, Grezes J. 2006. The power of simulation: Imagining one's own and other's behavior. Brain Research. 1079:4-14.

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. Cerebral 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. Magnetic Resonance in Medicine. 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. Psychological Review. 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. Nature Reviews Neuroscience. 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. Progress in Neurobiology. 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, Zahn R, de Oliveira-Souza R, Krueger F, Grafman J. 2005. The neural basis of human moral cognition. Nat Rev Neurosci. 6:799-809.
Moll J, Krueger F, Zahn R, Pardini M, de Oliveira-Souza R, Grafman J. 2006. Human fronto-mesolimbic networks guide decisions about charitable donation. P Natl Acad Sci USA. 103:15623-15628.
Moll J, de Oliveira-Souza R, Garrido GG, Bramati IE, Caparelli-Daquer EMA, Paiva MLMF, Zahn R, Grafman J. 2007a. The self as a moral agent: Linking the neural bases of social agency and moral sensitivity. Social Neuroscience. 2:336-352.
Moll J, De Oliveira-Souza R, Zahn R, Grafman J. 2007b. The cognitive neuroscience of moral emotions. In: Sinnott-Armstrong W, editor Morals in the brain: Emotion, disease and development Cambridge, MA: MIT Press.p
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. Journal of Affective Disorders. 71:19-27.
Piefke M, Weiss PH, Markowitsch HJ, Fink GR. 2005. Gender differences in the functional neuroanatomy of emotional episodic autobiographical memory. Human Brain Mapping. 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. Pers Soc Psychol Rev. 4:255-277.
Rorden C, Brett M. 2000. Stereotaxic display of brain lesions. Behavioural Neurology. 12:191-200.
Schwartz SH, Bilsky W. 1987. Toward a universal psychological structure of human-values. Journal of Personality and Social Psychology. 53:550-562.
Seymour B, Singer T, Dolan R. 2007. The neurobiology of punishment. Nature Reviews Neuroscience. 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. Biological 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, N.J.: 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. Nature Reviews Neuroscience. 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. Proceedings of the National Academy of Sciences (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.

Acknowledgements

This study was supported by NINDS intramural funding to JG and a German Academy of Natural Scientists Leopoldina Fellowship funded by the Federal Ministry of Education and Research (BMBF-LPD 9901/8-122) to RZ. GG was supported by the Brazilian Fundação de Amparo à Pesquisa do Estado de São Paulo grant 03/11794-6. JM was supported in part by the LABS-D'Or Hospital Network, Rio de Janeiro, Brazil. We thank Eric Wassermann for performing neurological exams, Kris Knutson and several SPM experts from the discussion list for imaging analysis advice.

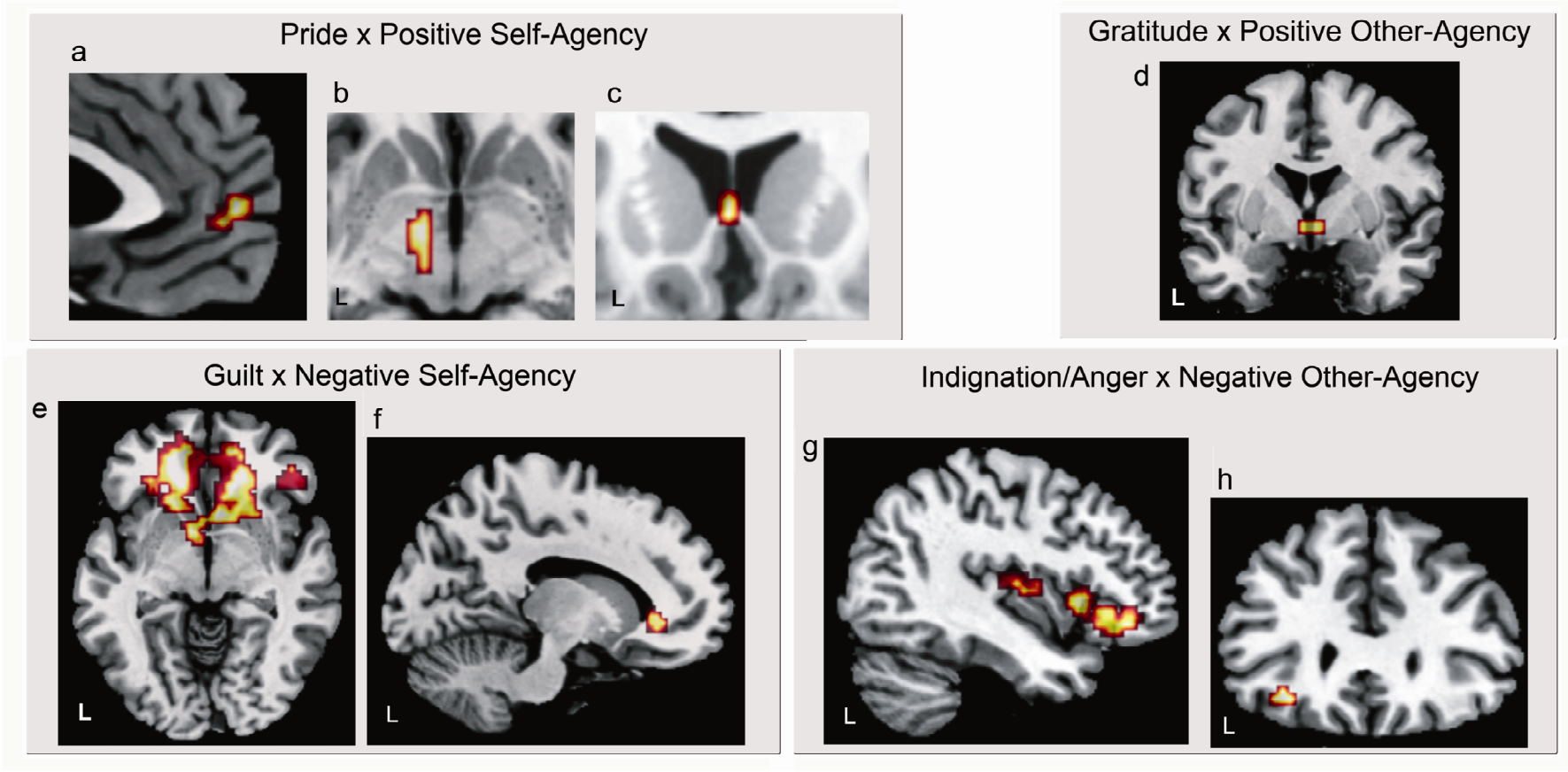


Fig 1

The partial effects for each CONDITION-SPECIFIC MORAL SENTIMENT compared with Fixation are displayed (see also Materials and Methods and Supplementary Tab 1). These are inclusively masked by two contrasts: 1) comparing the CONDITION-SPECIFIC MORAL SENTIMENT (e.g. Pride during POS_S-AG) vs. the condition with opposite valence and same agency role (e.g. Guilt during NEG_S-AG). 2) comparing vs. 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 apriori ROI analyses were carried out on the simple contrasts vs. Fixation without applying inclusive masking and all reported regions survived FWE-corrected P=.05 over apriori ROIs. All images are displayed at an uncorrected P=.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 vs. Pride and Indignation/Anger. g) Consistent group effect for Indignation/Anger, overlay of lateral OFC & insula ROIs. h) Consistent group effect for Indignation/Anger, whole brain. (see Supplementary Materials and Methods on definition of ROIs).

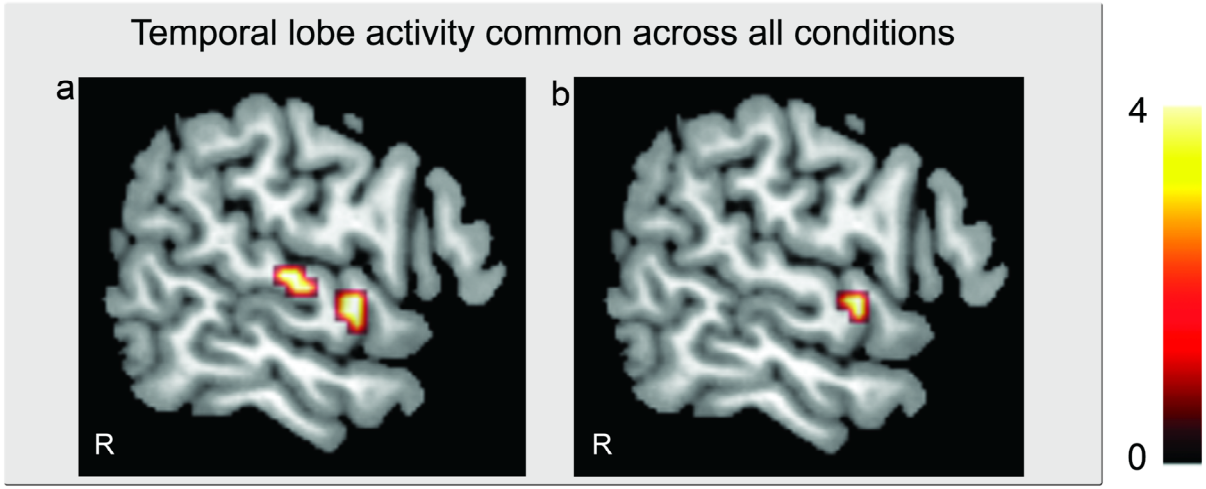


Fig 2
a) Partial effect of DESCRIPTIVENESS OF SOCIAL BEHAVIOR common across all conditions: right superior aTL (BA22, x=54, y=0, z=-3; T=4.53; FWE-corrected P over apriori ROI=.008) and right superior mid-posterior temporal gyrus (BA22, x=57, y=-18, z=6; T=4.6; FWE-corrected P over apriori ROI=.008). 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 apriori mask=.005). Activations are displayed at uncorrected P=.001, 5 voxels.

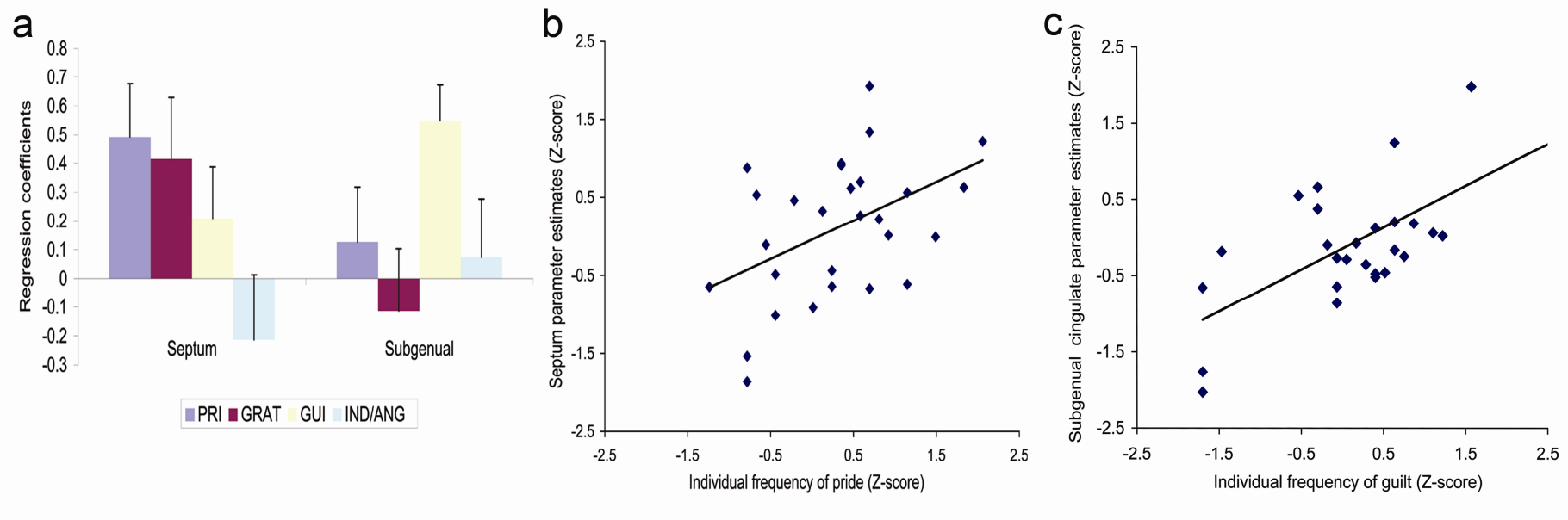


Fig 3

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 Tab 1, univariate ANOVA, SPSS14, outliers Z-score outside ±2.5 excluded): $F(107,7)=3.18$, $P=.03$. There was also a significant main effect of the moral sentiment covariate on the septal signal strength $F(114,1)=4.61$, $P=.03$. Unadjusted correlations for septal activity with pride for POS_S-AG: $R(29)=.46$, $P=.01$ (Trend for gratitude: $R(29)=.35$, $P=.06$, negative trend for indignation/anger: $R(28)=-.34$, $P=.08$, no significant correlations with guilt: $R(29)=.22$, $P=.25$). There was a significant interaction of condition with the moral sentiment covariate effect on the signal within the subgenual PFC: $F(112,3)=2.63$, $P=.05$. Subgenual PFC x NEG_S-AG_guilt: $R(28)=.66$, $P<.0001$ (no significant correlations with other moral sentiments: $P>.30$). There was a significant correlation of gratitude with hypothalamic activity ($R=.43$, $P=.02$). 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 $P=.05$). Thus the effects in the hypothalamus were not robust enough to survive adjustment on the secondary data analysis.

Supplementary Information

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

Roland Zahn, Jorge Moll, Mirella Paiva, Griselda Garrido, Frank Krueger, Edward D. Huey, Jordan Grafman

Supplementary Results

Influence of interindividual differences in strategy

According to our standardized questionnaire given after the scan, half of the subjects were frequently reminded of a specific episode or scene in their life during the fMRI task (N=14 of 29 with score > 4 on 1 to 7 Point Likert Scale for scene-likeness, adapted from (Piefke et al. 2005)), the other half showed low frequency of autobiographical episode retrieval (N=14 with score < 4). The same split was observed for visual imagery (14/29 with score > 4, 13/29 < 4). 3/29 subjects reported frequent vivid emotional experiences (>4 on 1 to 7 Point Likert scale). Visual imagery and scene-likeness were highly correlated (Spearman's rho=.82, P<.0001, 2-sided). There were also robust correlations between frequency of visual imagery with frequency of vivid emotional reactions (rho=.60, P=.001, 2-sided).

To rule out the theoretical possibility that observed effects within the superior anterior temporal lobe (aTL) common across conditions for DESCRIPTIVENESS OF SOCIAL BEHAVIOR was due to the subjects with a strong reliance on episodic autobiographical memory retrieval (highly correlated with imagery, see above), we

performed linear regressions with the peak voxel contrast estimate, removing the covariance contributed by interindividual differences in scene-likeness. This analysis demonstrated that there was no significant contribution of interindividual differences in strategy to the aTL activation and that the effect remained highly significant even after adjusting for the covariance of that effect (t[29]=3.39,P=.002 for the adjusted group mean, t[29]=-1.04,P=.31 for the regression coefficient of scene-likeness).

There were no correlations with interindividual differences in strategies and the frequency of the different moral sentiments (at P=.05), except for guilt showing significant positive correlations both with scene-likeness (Spearman's rho=.43, P=.02) and visual imagery (rho=.49, P=.006). To exclude, that individual differences in strategies could have influenced our correlations of regional fMRI signal with individual differences in moral sentiment frequency, we computed additional linear regressions (SPSS14) on peak voxel effects correcting for the variance contributed by scene-likeness and were able to confirm all individual difference effects reported for moral sentiments as independent of scene-likeness:

1) guilt on subgenual cingulate fMRI signal: partial effect of guilt: t[28]=5.35, P<.0001, partial effect of scene-likeness: t[28]=-1.25, P=.22

2) pride on septum fMRI signal: partial effect of pride: t[28]=2.57, P<.02, partial effect of scene-likeness: t[28]=-1.35,P=.19

3) gratitude on hypothalamus fMRI signal: partial effect of gratitude: t[28]= 3.81, P=.001, partial effect of scene-likeness: t[28]=-.74, P=.47

Discrepancy of whole brain and peak-voxel based analyses for gratitude

Higher individual frequency of gratitude was solely correlated with the hypothalamus on the whole brain analysis (Fig 1d, Supplementary Tab 1). The effect within the hypothalamus, however, was not significantly higher than in the other conditions on the confirmatory peak voxel analyses (statistics, see Figure 3). In order to understand this discrepancy, one should note, that the confirmatory peak voxel analysis may lead to different results than the whole brain based analysis. The most likely cause of this discrepancy is that for the whole brain based analysis we used pairwise comparisons between two individual moral sentiment effects and then masked different pairwise comparisons. For the confirmatory peak voxel analysis, however, we adjusted the effect of gratitude for all other moral sentiments (pride, indignation/anger, guilt) in one model which reduces the sensitivity of finding differences due to smaller adjusted correlations for gratitude. Therefore, the selectivity of hypothalamic activity for gratitude could not be as firmly established as the selectivity of septum and subgenual cingulate activity for pride and guilt.

Supplementary Materials and Methods

Subjects

To be able to compare our study population with our previous and subsequent study populations, we collected measures of self esteem and trait affective style before the fMRI experiment (Zahn et al. 2007) (Rosenberg Self Esteem Scale (Rosenberg 1989): mean=35.4±3, PANAS (Watson et al. 1988) positive affect score: mean=35.5±5.6; negative affect score: mean=14.9±4.6).

Psycholinguistic stimulus properties

Positive and negative concepts were defined according to social desirability norms (Z>0 => positive; Z<0 => negative, (Hampson et al. 1987)). We compared the two classes of concepts (positive, N=45 and negative, N=45), which were identical during the self-agency and other-agency condition, using independent samples t-tests for each of the 7 psycholinguistic variables (P=.05, word familiarity, Kucera Francis word frequency, word concreteness, word imageability from the MRC Psycholinguistic database (Coltheart 1981), number of syllables and DESCRIPTIVENESS OF SOCIAL BEHAVIOR and category-breadth from our previously published normative study (Zahn et al. 2007)). No difference emerged for word familiarity, Kucera Francis word frequency, concreteness, imageability or number of syllables.

Significant differences (t[88]>2.0, P<.05) were only detected for DESCRIPTIVENESS OF SOCIAL BEHAVIOR (negative > positive social concepts) and category-breadth (positive > negative social concepts), which had to be accounted for in the SPM model by using DESCRIPTIVENESS OF SOCIAL BEHAVIOR as a parametric covariate at the first level for each stimulus condition. Please note that category-breadth and DESCRIPTIVENESS OF SOCIAL BEHAVIOR were highly inversely correlated (Pearson's R=-.98, P = 0.0001) in our normative study (Zahn et al. 2007) so that correcting for one variable also adjusts for the variance on the other variable.

fMRI paradigm

Before the experiment, subjects were asked to name their best friend of same gender to whom they were not related by blood or marriage and to whom they entertained

a non-sexual relationship. Subjects saw verbal statements during event-related fMRI in which their best friend's and their own first nickname or name were used to indicate their agency role in a social behavior described by using an abstract social concept (e.g. 'Tom acts stingily towards Sam', see also Supplementary Fig 8). Subjects had to decide how they would feel about the described social behavior from their own perspective by pressing one of two response keys for 'pleasant' and 'unpleasant'. After the scan they performed ratings on a computer in a separate room with no experimenter present. Subjects had to choose a label (shame/embarrassment, guilt, indignation/anger, pride, gratitude, none/other) which best described their feeling associated with each statement. Subjects saw verbal statements or a visual pattern on a screen via a mirror system attached to the head coil. Stimuli (18-point font type) were back-projected onto a translucent screen placed at the feet of the participant with a magnetically shielded LCD video projector. The stimuli were presented in an event-related design with pseudorandom order of different stimulus types within each fMRI run and across the 3 runs. Visual stimulus presentation was controlled by ERTS (Experimental Run Time System, Berisoft Cooperation, Germany, http://www.erts.de). Each stimulus was displayed for 4 s, then a fixation asterisk appeared in the centre of the screen for a mean of 4.6 s (jittered from 2.6 to 6.6 s in 500 ms steps, see also Supplementary Fig 8). Subjects indicated whether each social behavior described by a verbal statement was pleasant or unpleasant from their own perspective by using the index or middle finger of their right hand to press a key. The order of runs and key/finger assignments to the related or unrelated decisions was counterbalanced across subjects. Subjects received a 5

min practice session on the actual task with a different set of stimuli to get familiarized with the experiment. Response times and responses were recorded for each trial.

Image analysis

Image analyses were performed using SPM5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5). The following pre-processing steps were applied: realignment and unwarping, slice timing correction, normalization $(3mm)^3$ voxel size), smoothing with a small kernel (FWHM = 6 mm) to be anatomically precise. Normalization was performed by first normalizing the 3D MPRAGE (3x3x3 mm voxel size) using the SPM5 T1- template and then applying the same transformation on the EPI images.

Estimated translation and rotation parameters were inspected and all subjects showed <3mm and 2 degrees of motion. As a basis function, we used the canonical hemodynamic response function (HRF) with time and dispersion derivative. On the first level of analysis we specified a general linear model (Friston et al. 1995) for each participant, which included the 5 trial types (POS_S-AG, POS_O-AG, NEG_S-AG, NEG_O-AG, Fixation=Null event). For each condition, DESCRIPTIVENESS OF SOCIAL BEHAVIOR of social concepts was modeled as a parametric predictor convolved with the HRF. Further the most frequently occurring moral sentiment (CONDITION-SPECIFIC MORAL SENTIMENT) was modeled as a categorical predictor (see Supplementary Fig 5) for the respective condition for each subject: POS_S-AG: pride, POS_O-AG: gratitude, NEG_S-AG: guilt, NEG_O-AG: indignation/anger. Lastly, self-distinctiveness of concepts was modeled as a parametric predictor of no

interest. Rated closeness towards the subjects' best friend (1 to 7 Point Likert Scale, ratings ranged from 5 to 7) significantly decreased the mean self-distinctiveness of social values (One-way ANOVA $F[2,61]=3.69$, $P=.03$). This indicates that there is a relationship between self-distinctiveness of social values and the overall feeling of proximity even towards a very familiar other.

The variables (DESCRIPTIVENESS OF SOCIAL BEHAVIOR, CONDITION-SPECIFIC MORAL SENTIMENTS and SELF-DISTINCTIVENESS) used in our model did not show strong correlations among each other ($R<.25$) and there was no association of increased response time (RT) with any of the predictors. Moral sentiment trials were instead associated with decreases in RT (Pearson's R between -.44 and .-65, $P<.003$, 2-tailed). The decreasing effect on RT for moral sentiments (pride, gratitude, guilt, indignation/anger) was equal across conditions. This was tested by looking at the interaction of moral sentiment effect (% of subjects with the moral sentiment of interest per stimulus, coding all moral sentiments of interest into one variable) and condition (POS_S-AG, POS_O-AG, NEG_S-AG, NEG_O-AG) on the mean RT per stimulus ($N=180$) using a univariate ANOVA within the General Linear model in SPSS14. Whereas there was a highly significant decrease of RT associated with moral sentiments overall ($F[1,172]=66.48$, $P<.0001$), there was no interaction of condition and moral sentiment effect on RT ($F[3,172]=.21$, $P=.89$). DESCRIPTIVENESS OF SOCIAL BEHAVIOR was not associated with RT increases or decreases ($R=.08$, $P=.31$, 2-tailed).

A measure of multivariate distance of statements (Mahalanobis) on all these variables was computed for each condition ($N=45$) to exclude multivariate outliers, all values were within the 99.5 percentile (Stevens 1996). Univariate outliers were excluded

by Z-transformation of per stimulus scores over 180 stimuli for DESCRIPTIVENESS OF SOCIAL BEHAVIOR, % of subjects with moral sentiment of interest for that condition and self-distinctiveness. All Z-values were within the expected range for normally distributed data (maximally 1% of values between 2.5 and 3 or between -3 and -2.5, no value outside +/-3)

To derive subject-specific covariates for the second level random-effect analyses, we computed Z-scores for each subject compared to the total sample of N=64 of the normative study in order to remove the effects of differences in the overall frequency of occurrence of each moral sentiment in each condition. These Z-scores therefore indicate the individual percentage of trials on which a given moral sentiment was experienced in the respective condition by a participant with reference to the normative sample.

Definition of apriori regions of interest and anatomical localization

All analyses were performed at the whole brain level and within predefined bilateral apriori regions of interest: anterior temporal lobe (aTL), posterior superior temporal sulcus/temporo-parietal junction (pSTS/TPJ), dorsolateral PFC, ventromedial PFC, lateral orbitofrontal cortex (OFC), dorsomedial PFC, primary and supplementary motor cortex (Motor), insula, amygdala, basal ganglia, septum, hypothalamus, ventral tegmental area (VTA).

The aTL ROI was identical to the one used in our previous study (Zahn et al. 2007). The ROI was created using bilateral Brodmann (38, 22, 21) maps from the WFU Pickatlas (Maldjian et al. 2003) integrated in SPM5. The original maps were cropped to exclude tissue posterior to MNI y coordinate = -10 . The following ROIs were taken from

the Automatic Anatomical Labeling atlas (AAL, (Tzourio-Mazoyer et al. 2002)) as implemented in the WFU Pickatlas (Maldjian et al. 2003): Insula ROI=AAL insula, Motor ROI= combined AAL precentral and supplementary motor area, amygdala ROI=AAL amygdala, basal ganglia ROI=combined AAL striatum and pallidum. The lateral OFC ROI was created by combining bilateral AAL ROIs for the orbital parts of the inferior, middle and superior frontal gyri and the olfactory gyri. Then, based on the meta-analysis by Kringelbach & Rolls (Kringelbach and Rolls 2004) we defined the lateral parts of the OFC as lateral to MNI x=20 by importing the AAL masks into MRIcron (Rorden and Brett 2000) and cropping the medial parts. The ventromedial PFC ROI was created by combining bilateral AAL ROIs medial superior frontal, anterior cingulate, orbital parts of inferior, middle and superior frontal gyri, olfactory cortex and gyrus rectus cropped laterally to only include cortex medially to MNI x=21 and dorsally to only include cortex ventrally to MNI z<1 (genu of the corpus callosum). The posterior borders of the AAL region were manually extended to include the subcallosal area which had been partly not included on the original mask.

The dorsomedial PFC ROI was created by combining bilateral AAL ROIs for medial superior frontal, anterior and mid cingulate and cropped to include only cortex dorsally to MNI z=1 . The dorsolateral PFC ROI was formed by combining bilateral AAL ROIs for superior, middle frontal gyri and pars triangularis of the inferior frontal gyrus (none of these regions reached more ventrally than MNI z=1).

The pSTS/TPJ ROI was created by combining bilateral AAL ROIs for angular, supramarginal, inferior parietal and superior and middle temporal gyri and cropped to only include cortex posterior to MNI y=-12. This was done to include all superior

temporal cortex falling outside of our aTL ROI and to include all activations reported for biological motion and social perception (based on the systematic review of Allison et al. (Allison et al. 2000) on the STS region with the most anterior activation found for visual biological motion at y=-20).

The septum ROI was delineated according to microscopic sections ((Nieuwenhuys et al. 1978), pp 65/66) showing the septal nuclei at coronal sections through the anterior commissure and anterior to the optic chiasma. The septal nuclei are contained within the left and right hemispheric and inferior parts of the septum which converge superiorly into the midline part separating left and right parts of the third ventricle. The septum is in the midline superior to the anterior ventral striata and posterior to the area subcallosa (i.e. subgenual region). We used these anatomical landmarks to manually delineate the septum on the standard template using MRIcron (Rorden and Brett 2000). The ascending parts of the septum were delineated from MNI y=0 to MNI y=14 .

The hypothalamic region was defined using the Talairach atlas (Talairach and Tournoux 1988) where the hypothalamus is depicted between 3 and 5 mm on the x axis, between +4 and -8 on the y-axis and between -10 and -5 on the z axis. On the Nieuwenhuys atlas (Nieuwenhuys et al. 1978) the hypothalamus is medially limited by the fornix and the wall of the third ventricle, anteriorly by the lamina terminalis which runs inferiorly from the anterior commissure down to the chiasma opticum. The corpora mamillaria are a landmark posterior to the hypothalamus. The Talairach atlas shows the nucleus accumbens and ventral pallidum as superior-lateral border landmarks at y=+4 and z coordinates of -10 to -12. We defined the ROI conservatively not including the very

posterior part which cannot be distinguished from adjacent nuclei (e.g. nucleus subthalamicus), in Talairach space: x: 2 to 8, y: -1 to -5, z: -4 to -12 corresponding to MNI: x: 2 to 8, y: -1 to -5, z: -3 to -15 (using Matthew Brett's formula, http://www.mrc-cbu.cam.ac.uk/Imaging/Common/mnispace.shtml). On MRIcron using the standard template in MNI space we adapted these borders according to the anatomical landmarks (thalamus, ventral pallidum, ventral striatum) with cubic regions drawn on coronal slices and checked on 3 dimensions (MNI): y: -1 to -3, x: +/-5 to 3, z: -9 to -12 and y: -4 to -5, x: +/-2 to 8, z: -5 to -11. Smoothed VTA and hypothalamus ROIs were overlapping but showed different centers. Therefore to localize activations we used all overlapping ROIs and reported the localization as within the ROI in which most voxels survived the ROI analysis, further we identified the activation peak with reference to anatomical atlas landmarks.

The Ventral Tegmental Area (VTA) can only be precisely defined microscopically within the midbrain (Nieuwenhuys atlas, p. 74). It can be detected on MRI, however, by using anatomical landmarks: the VTA is dorsally adjacent to the mamillary body, ventrally to the red nucleus, medially to the substantia nigra, having the midline as its medial border. On the Talairach atlas, axial slice z=-8 most closely matched the Nieuwenhuys slice cutting through corpora mamillaria, red nucleus and substantia nigra, coronal slices at Tal y=-8 to y=-12 show corpora mamillaria at z=-8 to -10 adjacent to the midline.

On axial Talaraich slice z =-8 we therefore defined the VTA as the area between mamillary bodies (ventral border), substantia nigra (lateral border), red nucleus (dorsal border, extension in z-direction until -4) and derived the following coordinates:

Talairach x: 5 to -5, y: -7 to -12, z: -8 to -10 => MNI x: 5 to -5, y: -7 to -12, z: -6 to -8. We validated those coordinates by comparing the anatomical landmarks within these borders on the Colin standard template in MNI-space used within MRIcron.

Localization of areas was determined by using anatomical landmarks (Mai et al. 2004; Nieuwenhuys et al. 1978; Talairach and Tournoux 1988) and by looking at activations in original MNI space projected onto a standard MNI template. In addition Talairach transformed coordinates (using Matthew Brett's formula, http://www.mrc-cbu.cam.ac.uk/Imaging/Common /mnispace.shtml) were used to identify corresponding Brodmann Areas on the Talairach atlas (Talairach and Tournoux 1988) using Talairach Daemon software version 2 (http://ric.uthscsa.edu/projects/talairachdaemon.html).

Supplementary References

Allison T, Puce A, McCarthy G. 2000. Social perception from visual cues: Role of the sts region. Trends in Cognitive Sciences. 4:267-278.

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, Human experimental psychology. 33:497-505.

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

Hampson S, Goldberg L, John O. 1987. Category-breadth and social-desirability values for 573 personality terms. European journal of personality. 1:241-258.

Kringelbach ML, Rolls ET. 2004. The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. Progress in Neurobiology. 72:341-372.

Mai J, Assheuer J, Paxinos G.2004. Atlas of the human brain. Amsterdam / Boston: Elsevier Academic Press.

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.

Nieuwenhuys R, Voogd J, Van Huijzen C.1978. The human central nervous system. Berlin, Heidelberg, New York: Springer.

Piefke M, Weiss PH, Markowitsch HJ, Fink GR. 2005. Gender differences in the functional neuroanatomy of emotional episodic autobiographical memory. Human Brain Mapping. 24:313-324.

Rorden C, Brett M. 2000. Stereotaxic display of brain lesions. Behavioural Neurology. 12:191-200.

Rosenberg M.1989. Society and the adolescent self-image. Middleton, CT: Wesley University Press.

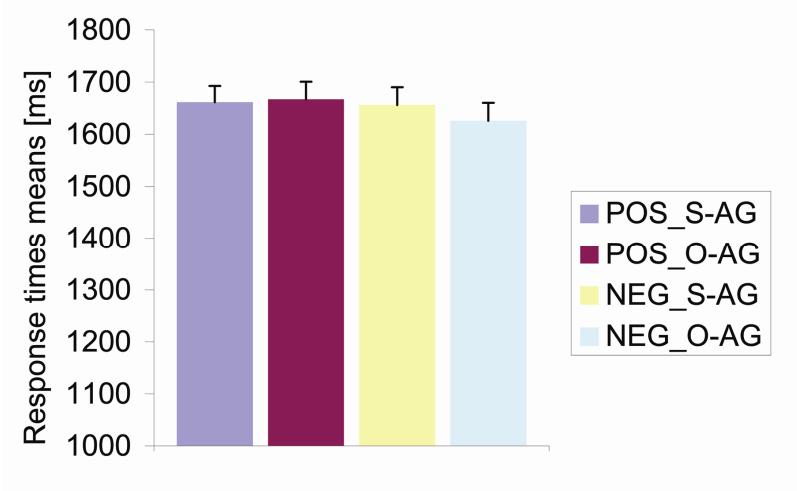
Stevens J.1996. Applied multivariate statistics for the social sciences. Mahwah, N.J.: Lawrence Erlbaum Associates.

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

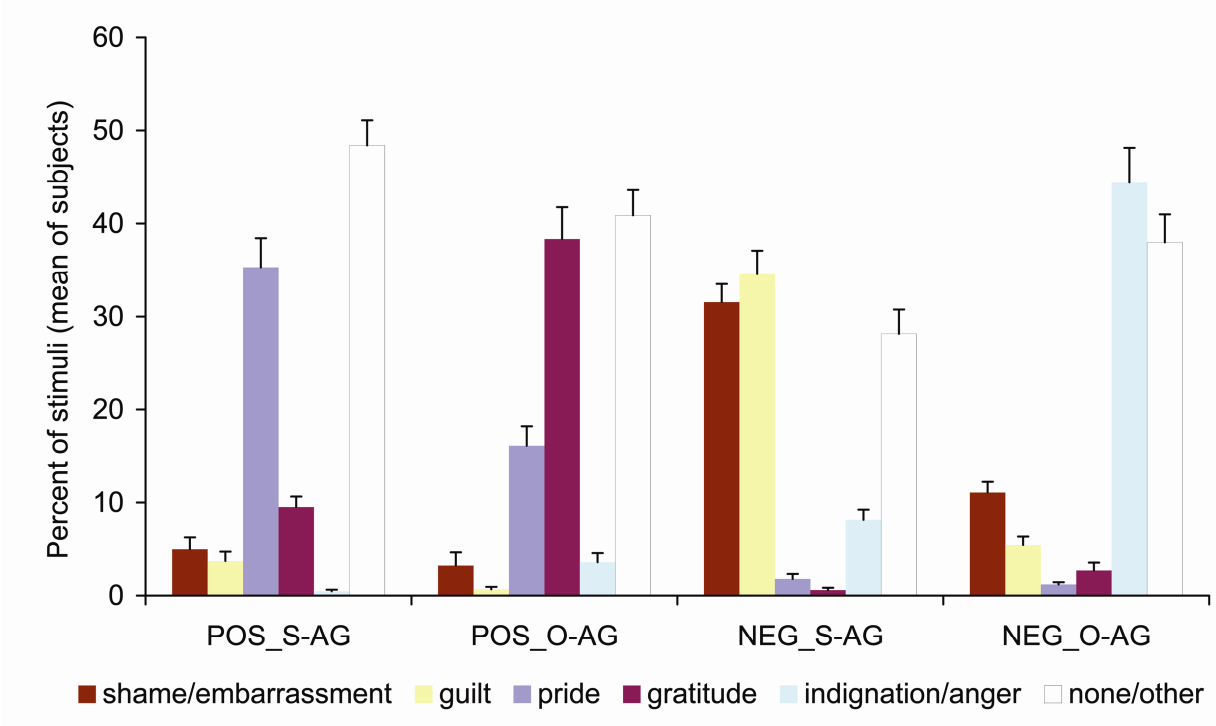
Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. 2002. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. Neuroimage. 15:273-289.

Watson D, Clark LA, Tellegen A. 1988. Development and validation of brief measures of positive and negative affect: The panas scales. J Pers Soc Psychol. 54:1063-1070.

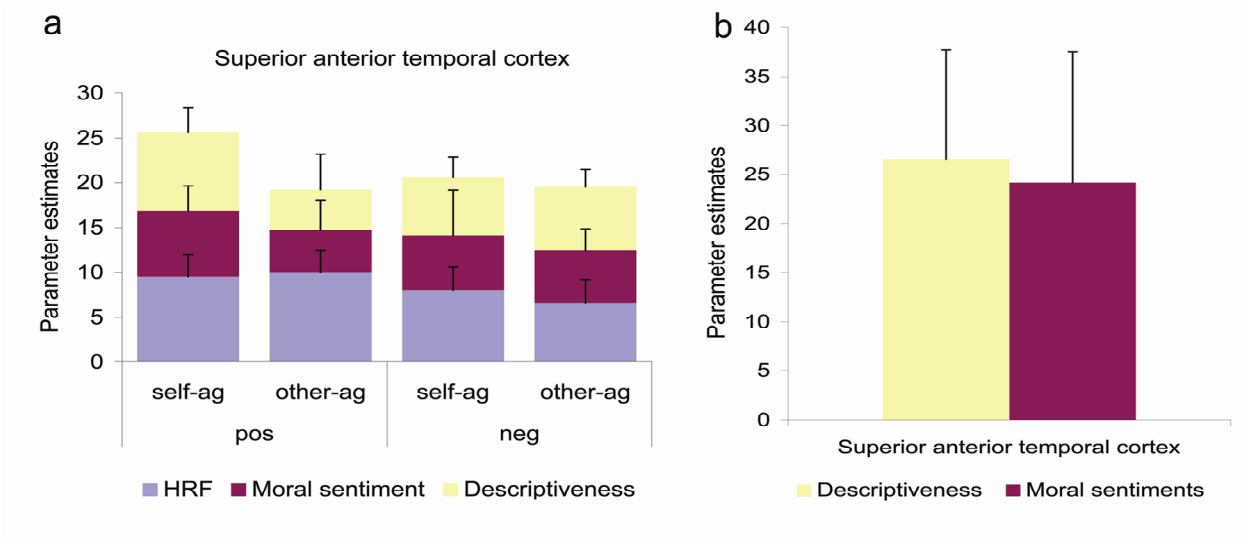
Zahn R, Moll J, Garrido G, Krueger F, Huey ED, Grafman J. 2007. Social concepts are represented in the superior anterior temporal cortex. Proceedings of the National Academy of Sciences (USA). 104:6430 - 6435.



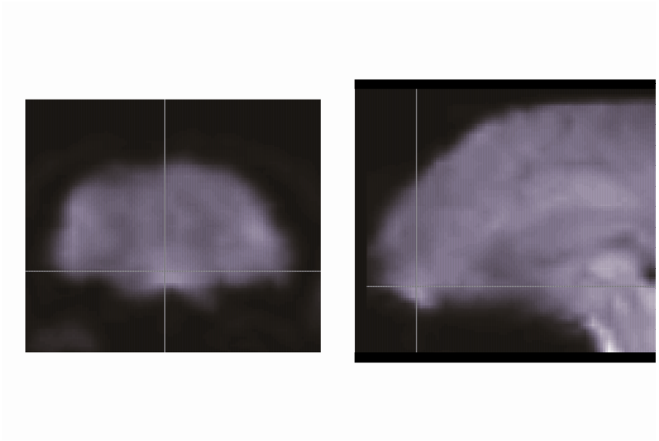
Supplementary Figure 4
Means and standard errors for response times. There was no significant difference in mean response times over the 29 subjects during fMRI between the 4 experimental conditions, N=45 stimuli per condition (one-way ANOVA, $F[3,176]=.47$, $P=.70$). The percentage of subjects responding with 'pleasant' and 'unpleasant' was not influenced by agency (agency effect on positive conditions: N=90 stimuli, Mann-Whitney $U=980$, asymptotic $P=.79$, 2-tailed; negative conditions: N=90 stimuli, Mann-Whitney-$U=887$,$P=.31$). The number of missed responses was low ($m=1.84\pm2.35\%$ of subjects per stimulus) and not significantly different across conditions (Kruskal Wallis Test $\chi[3]=6.64$, asymptotic $P=.08$).



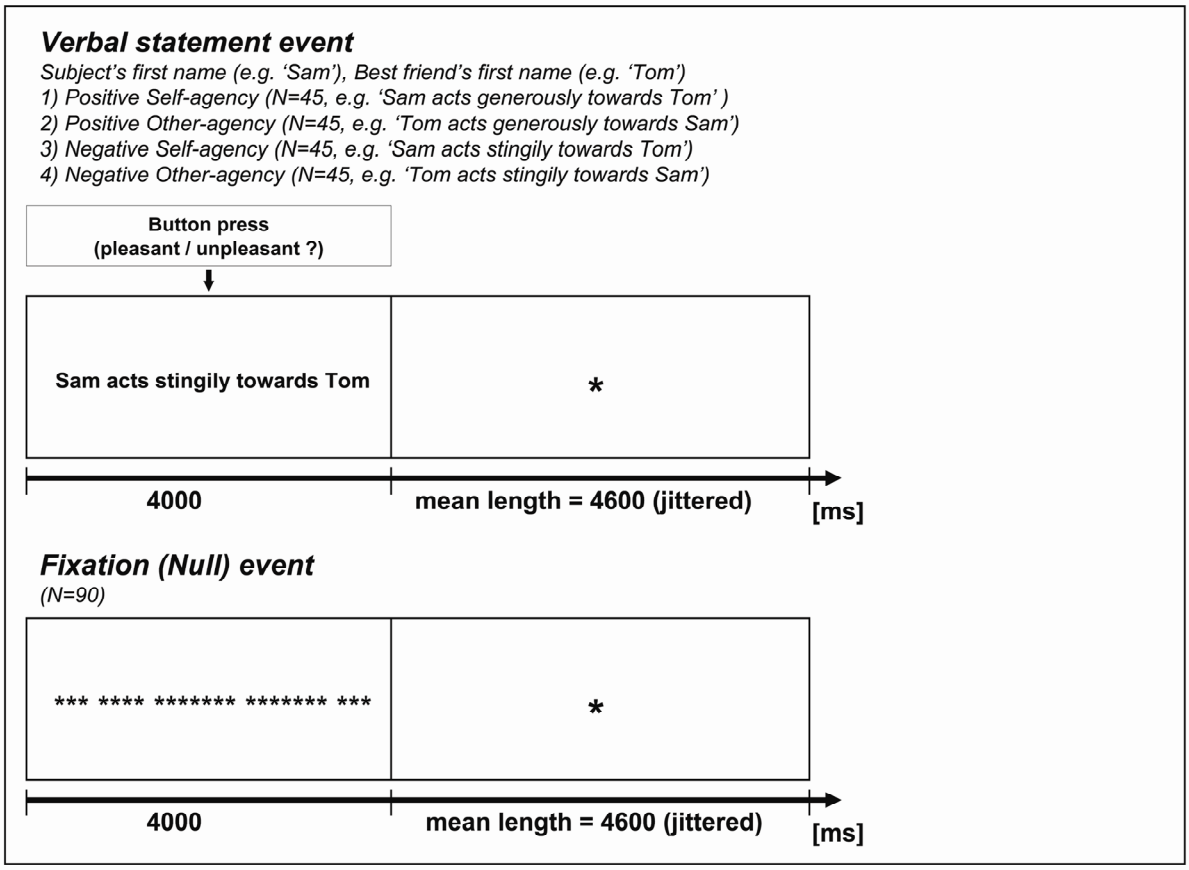
Supplementary Figure 5
Frequency of moral sentiments (mean percentages and standard errors over 45 stimuli per condition for N=64 of the normative study). Guilt was slightly more frequent than shame/embarrassment and therefore chosen as a sentiment of interest. There were no correlations between guilt and shame/embarrassment (Pearson R=-.194, P=.20). The most frequent moral sentiment in each condition was chosen as CONDITION-SPECIFIC MORAL SENTIMENT for this study. Paired samples t-tests were used to compare frequencies of CONDITION-SPECIFIC MORAL SENTIMENTs between self-agency and other-agency conditions within each valence category (positive/negative) and demonstrated highly significant effects of agency within each valence category (excluding main effects of valence to explain the frequency of CONDITION-SPECIFIC MORAL SENTIMENTs): pride in POS_S-AG vs. POS_O-AG (t[44]=8.94, P<.0001), gratitude in POS_O-AG vs. POS_S-AG (t[44]=10.55, P<.0001), guilt in NEG_S-AG vs. NEG_O-AG (t[44]=10.30, P<.0001), indignation/anger in NEG_O-AG vs. NEG_S-AG (t[44]=10.80, P<.0001). There were no differences in rated familiarity from personal experience or rated pleasantness/unpleasantness (i.e. valence) between self- and other-agency conditions (independent samples t-test, 2-tailed: familiarity: t[178]=02, P=.98, pleasantness/unpleasantness: t[178]=-.35, P=.73). As to be expected, rated pleasantness/ unpleasantness was, however, significantly higher in the positive than the negative conditions t[178]=23.31, P<.0001).



Supplementary Figure 6
Partial effects of HRF, DESCRIPTIVENESS OF SOCIAL BEHAVIOR and CONDITION-SPECIFIC MORAL SENTIMENT within superior aTL peak voxel (see Fig 2b) for all 4 conditions with no differences across conditions (Repeated measures ANOVA, SPSS14: main effects and interaction of valence and agency: $F[28,1]<1.4$, $P>.25$) and significant difference from zero within each condition tested separately ($T[29]>2.6$, $P<.01$). b) Parameter estimates for the sum of partial effects of DESCRIPTIVENESS OF SOCIAL BEHAVIOR over all conditions and sum of partial effects of CONDITION-SPECIFIC MORAL SENTIMENT over all conditions within the superior aTL peak coordinate (from Fig 2b) are displayed. Both effects are significantly different from zero (DESCRIPTIVENESS OF SOCIAL BEHAVIOR: $T[28]=3.42$,$P=.002$; CONDITION-SPECIFIC MORAL SENTIMENT: $T[28]=2.45$,$P=.02$) and there was no significant difference between both partial effects ($T[28]=.30$, $P=.77$). Displayed are means and standard errors.

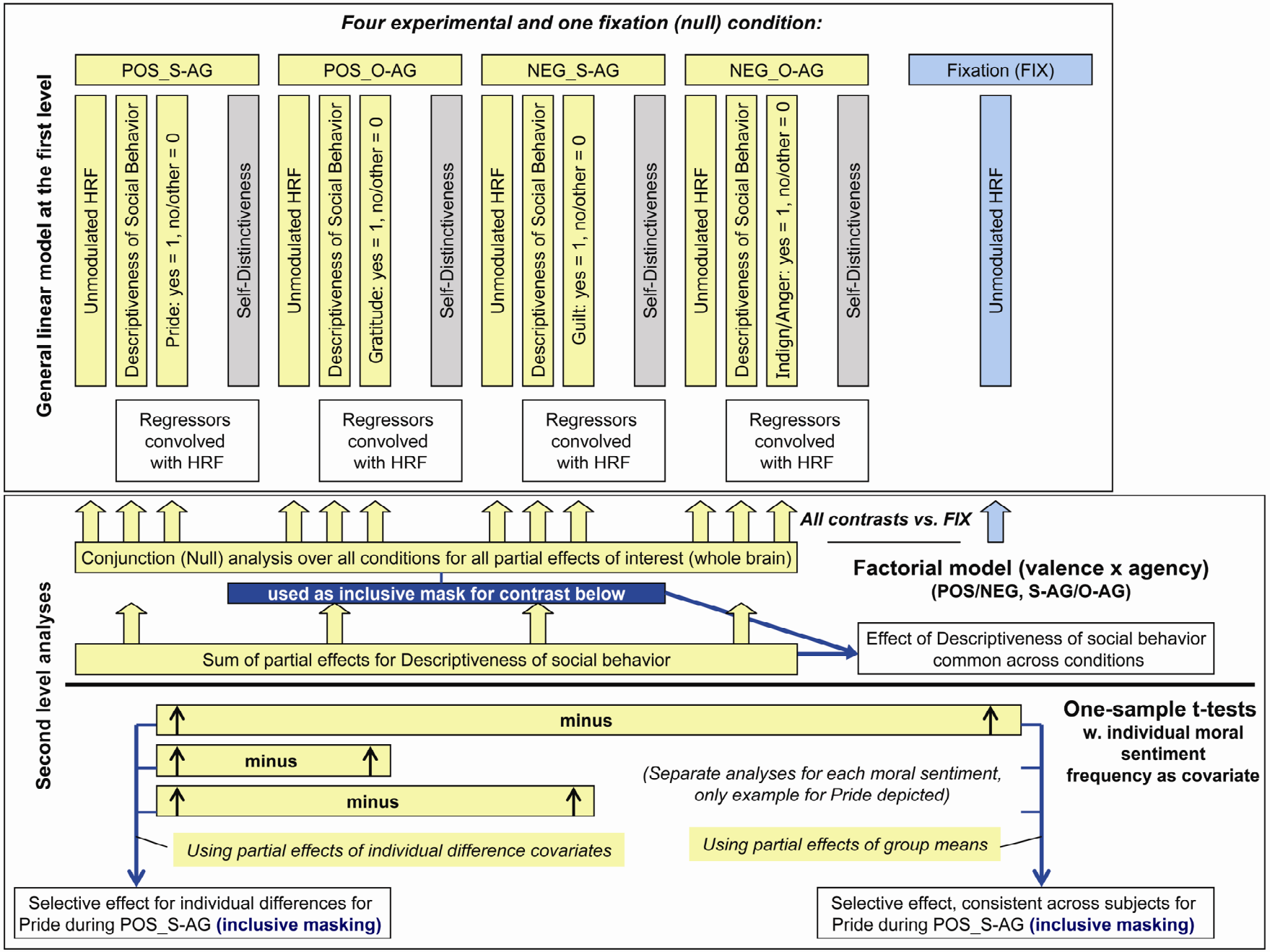


Supplementary Figure 7
Coronal, sagittal and axial slices through the anterior ventromedial PFC of a normalized echoplanar image from a representative participant (x=1, y=56, z=-11) shows full coverage of most of the ventral frontal lobes. Ventromedial PFC including BA25, ventral BA24, ventral BA32, ventromedial BA 10 were fully covered. Susceptibility artefacts in ventral frontal lobes were minimized by using reduced slice thickness (Bodurka et al. 2007) and manual shimming. Unavoidable and typical signal loss was observed within anterior medial BA11 due to susceptibility artefacts near the nasal cavities. All subjects included in the study had coverage of anterior ventromedial cortex covering cortex ventrally up to z=-6 in all subjects (at y=50). The lateral OFC was almost fully covered with typical signal loss within the most lateral ventral parts (coverage reliably included cortex dorsally to z=-19 in all subjects). aTL and all other ROIs were fully covered. One subject with ventromedial signal loss exceeding the normal limits had to be excluded prior to the statistical analysis. All normalized EPI images were visually inspected and critical regions were examined using above coordinates as criteria of coverage. One subject was excluded prior to statistical analysis because of signal drop out within predefined critical regions (aTL BA38/22, 21, 20; ventromedial PFC, BA 11,25,24,32; lateral OFC, BA11/47; and frontopolar cortex, BA10).



Supplementary Figure 8

The fMRI paradigm. Subjects saw verbal statements containing their own and their best friend's first name (e.g. 'Tom acts stingily towards Sam') or a visual pattern during fMRI. Each stimulus was displayed for 4 s, then a fixation asterisk appeared on the screen for a mean of 4.6 s (jittered interval). Subjects indicated whether each social behavior described by the statement was pleasant or unpleasant from their own perspective by pressing a key during the display period.



Supplementary Figure 9
Schematic outline of the most important image analyses using SPM5. Details see Methods and Materials.

Supplementary Table 1 Categorical effects of moral sentiments and interindividual differences

Contrast	Hemi-sphere	Area	x	y	z	T-score
Pride x POS S-AG: interindividually consistent	L	Anterior ventromedial prefrontal (BA10)	-9	54	-3	4.54#
	L	Ventral tegmental area / posterior hypothalamus	-9	-9	-6	4.26#
	L	Parahippocampal gyrus (BA30)	-9	-48	3	5.82*
Pride x POS S-AG: interindividual difference	L&R	Septum	0	15	6	3.18^
Gratitude x POS O-AG: interindividual difference	R & L	Hypothalamus / ventral tegmental area	3	-3	-3	4.04#
Guilt x NEG S-AG: interindividual difference	L	Subgenual cingulate gyrus/sulcus BA32	-15	36	-6	5.48*#
	L	Anterior ventromedial prefrontal (BA10)	-21	51	-3	3.67*#
	R > L	Ventral anterior cingulate (BA32)	6	42	-3	3.19*#
Indignation/anger x NEG O-AG: interindividually consistent	L	Lateral orbitofrontal (BA47)	-30	30	-12	4.81#
	L	Insula	-36	15	3	3.94#
	L	Dorsolateral inferior frontal (BA9)	-45	6	30	4.16#
	R	Medial premotor cortex (BA6)	3	-3	54	5.00#

All analyses were performed using an uncorrected P=.005, 4 voxels, only regions surviving a Family-Wise-Error (FWE)-corrected threshold of P=.05 over apriori defined anatomical ROI volumes are reported. Non-predicted regions were reported if they survived whole-brain correction for multiple comparisons (FWE) at P=05. Only regions which survived inclusive masking with two complex contrasts are reported (P=.05, 5 voxels uncorrected): 1) condition-specific moral sentiment effect vs. same agency & opposite valence condition 2) condition-specific moral sentiment effect vs. same valence & opposite agency-condition. ^ Within the apriori defined septum ROI (407 voxels) we used the more sensitive set-level inference: corrected P=.009, FWE-corrected P=.11. Coordinates are in Montreal Neurological Institute Standard Space (MNI). Areas surviving FWE-corrected $P \leq .05$ over the whole brain volume are marked with *, areas surviving FWE-corrected $P \leq .05$ over apriori ROI are marked with #. Negative effects of moral sentiments were also tested and no significant clusters for interindividual differences or consistent group effects emerged. No significant clusters were found for Gratitude x POS O-AG (interindividually consistent), Guilt x NEG S-AG (interindividually consistent) and Indignation/anger x NEG O-AG (interindividual difference).

Supplementary Table 2 Partial effect of descriptiveness of social behavior common across conditions

Hemi-sphere	Area	Partial effect of descriptiveness of social behavior over all conditions in areas common across conditions			
		x	y	z	T-score
L	Orbitofrontal cortex (BA11/47)	-30	33	-12	4.68*#
L	Anterior dorsolateral prefrontal cortex (BA10)	-24	57	9	5.54*#
R	Anterior superior temporal gyrus (BA22)	54	0	-3	4.53#
R	Posterior superior temporal gyrus (BA22)	57	-18	6	4.6#
L	Lingual gyrus (BA19)	-12	-51	0	5.85*
L	Posterior thalamus	-18	-27	0	5.49*
R	Anterior cerebellum	33	-57	-27	7.25*
R	Posterior thalamus/ retrosplenial cortex (BA30)	21	-30	-3	5.79*
R	Paracentral lobule (BA5)	6	-42	60	5.06*

Partial effects of descriptiveness of social behavior over all conditions were inclusively masked by a conjunction null analysis over all conditions and partial effects at P=.001, 5 voxels and exclusively masked by F-tests for main effects and interactions of valence and agency at P=.05, uncorrected to rigorously exclude regions with differences across conditions. All analyses were performed using an uncorrected P=.001, 5 voxels, only regions surviving a Family-Wise-Error (FWE)-corrected threshold of P=.05 over apriori defined anatomical ROI volumes are reported. Non-predicted regions were reported if they survived whole-brain correction for multiple comparisons (FWE) at P=.05. Coordinates are in Montreal Neurological Institute Standard Space (MNI). Areas surviving FWE-corrected P≤.05 over the whole brain volume are marked with *, areas surviving FWE-corrected P≤.05 over apriori ROI are marked with #.