doi:10.1093/scan/nsq013 SCAN (2010) 5, 414–423

Unfair? It depends: Neural correlates of fairness in social context

Berna Güroğlu,[1,2] Wouter van den Bos,[1,2] Serge A. R. B. Rombouts,[1,2,3] and Eveline A. Crone[1,2]
[1]Institute of Psychology, Leiden University, [2]Leiden Institute for Brain and Cognition and [3]Department of Radiology, Leiden University Medical Center, the Netherlands

Fairness is a key concept in social interactions and is influenced by intentionality considerations. In this functional magnetic resonance imaging study, we investigated the neural correlates of fairness by focusing on responder behavior to unfair offers in an Ultimatum Game paradigm with conditions that differed in their intentionality constraints. Brain activity underlying rejection *vs* acceptance of unfair offers appeared highly dependent on intentionality. Rejection of unfair offers when the proposer had no-alternative as well as acceptance of offers when the proposer had a fair- or hyperfair-alternative was associated with activation in a network of regions including the insula and the dorsal medial prefrontal cortex. These activations were interpreted as neural responses to norm violations because they were mostly involved when behavior was inconsistent with socially accepted behavior patterns. Rejection of unfair offers in the no-alternative condition further resulted in activity in the anterior medial prefrontal cortex and the temporoparietal junction, which was interpreted in terms of higher moral mentalizing demands required in social decision-making when rejection could not be readily justified. Together, results highlight the significance of intentionality considerations in fairness-related social decision-making processes.

Keywords: decision-making; fairness; intentionality; insula; social

INTRODUCTION

In social interactions fairness emerges as a key concept, in such a way that individuals are not solely motivated by self-interest but also by self *vs* other comparisons (Fehr and Fischbacher, 2003). In the recent years, behavioral studies based on economic games have clearly demonstrated that individuals are not purely rational beings aiming to maximize self-gain but also care about their relative benefits (Camerer, 2003). Other-regarding preferences and its comparison with self-gain have been captured well by the Ultimatum Game (UG). In the original version (Güth *et al.*, 1982) a player (responder) is faced with a monetary offer from another player (proposer) who is asked to divide a stake between the two players. Whereas accepting the offer leads to the suggested division of the stake, rejection of an offer results in both players going empty-handed. According to the self-gain maximization principle, responders are expected to accept every offer above zero. However, when faced with unfair divisions of the stake, responders often reject offers, thereby preferring that both players receive nothing (Güth *et al.*, 1982). This subjective comparative component of decision-making is associated with *fairness* judgments and entails the comparison between maximizing self-gain and relating self-gain to outcomes for others. The latter process is thought to bring additional social and emotional aspects into the decision-making process.

Within the decision-making process involving fairness considerations, we can identify several automatically intertwined steps that prove difficult to disentangle. Besides the comparative component of self *vs* other gain, individuals automatically integrate context-related information into the decision-making process to further evaluate the comparison at hand. One such context dependent information that highly influences fairness judgments is intentionality, that is, perceptions of fairness are influenced by the intentions of the interaction partner (Fehr and Schmidt, 1999; Lee, 2008). A seemingly unfair act might evoke less negative affect if one believes that it was not done intentionally. Within the UG paradigm, the decision of the responder to accept or reject an offer reflects the outcome of the decision-making process which entails both perception (i.e. how an offer is perceived) as well as evaluation (i.e. how the offer has been evaluated).

Neuroscientific studies have identified several brain regions that are involved in self-other comparisons when individuals play the UG. Fairness has been related to activation in brain regions which have previously been associated with negative and positive affect, including the insula, the ventral striatum, amygdala and the orbitofrontal cortex (OFC). Sanfey *et al.* (2003) reported that perception of unfair offers resulted in activation of bilateral insula, dorsolateral prefrontal cortex (DLPFC) and the anterior cingulate cortex (ACC) (Sanfey *et al.*, 2003). In contrast, Tabibnia

Received 17 July 2009; Accepted 13 January 2010
Advance Access publication 28 March 2010
The authors would like to thank Eric van Dijk for his advice and Félice van Nunspeet, Mariska Okkinga, and Bianca van den Bulk for their assistance in the data collection.
This research was supported by a VIDI grant from the Netherlands Organisation for Scientific Research (NWO) to the author E.A. Crone.
Correspondence should be addressed to Berna Güroğlu, Department of Developmental Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands. E-mail: bguroglu@fsw.leidenuniv.nl.

et al. (2008) reported no differential activation following low-fairness offers compared to high-fairness offers (Tabibnia *et al.*, 2008). Fair offers, in contrast, resulted in more activation in the ventral striatum, amygdala and OFC relative to unfair offers (Tabibnia *et al.*, 2008), which is thought to indicate that fairness is a rewarding social stimulus. Indeed, individuals reported higher levels of positive affect (i.e. happiness) when they received fair offers than when they received unfair offers (Tabibnia *et al.*, 2008), whereas skin conductance activity is higher for unfair than fair offers (van't Wout *et al.*, 2006).

Insula activation associated with perception of 'unfairness' has been interpreted as negative emotional reaction to unfair offers. This interpretation is based on prior studies reporting insula involvement in negative or arousing emotional states such as fear, disgust, or anger, as well as pain and autonomic arousal (Phan *et al.*, 2004) and is further supported by higher levels of anterior insula activation during rejection of unfair offers (Sanfey *et al.,* 2003; Tabibnia *et al.*, 2008). Alternatively, insula involvement could also be associated with error signals associated with norm violations (i.e. receiving an unfair offer) (Montague and Lohrenz, 2007; King-Casas *et al.*, 2008). This perspective is supported by findings that extend the role of the anterior insula beyond representation of emotional and physiological states to prediction errors related to expected states and uncertainty (Singer *et al.*, 2009), as well as by findings that relate insula activity to social norm compliance (Spitzer *et al.*, 2007). The first step in the decision-making process entails evaluation of internal (e.g. fairness norms or self-interest goal) and external states (e.g. context information such as intentionality) as well as evaluation of possible courses of action (Rangel *et al.*, 2008). Negative emotionality associated with decision-making related to unfairness is possibly registered along with an error signal from internal states (e.g. discrepancy between fairness expectation and unfair offer) suggesting the individual performs behavior (e.g. rejection of unfair offer) that restores the norm (e.g. by punishing unfair proposer) and/or internal state (e.g. negative emotionality).

Another brain region that has been shown to play a role in decision-making during the UG paradigm is the DLPFC, which is commonly associated with top–down executive functions in controlling impulses in decision-making to accept offers (Rilling *et al.*, 2008). For example, several transcranial magnetic stimulation studies that interfere with rDLPFC activity point out that diminished executive control is related to higher acceptance of unfair offers, possibly suggesting DLPFC involvement in control of selfish impulses when faced with an unfair offer (Wout *et al.*, 2005; Knoch *et al.*, 2006). Similarly, norm compliance through control of prepotent behavior related to self-gain is found to involve DLPFC activation (Spitzer *et al.*, 2007).

These studies, however, have not examined context related information, such as intentionality of offers, in examining neural correlates of fairness judgments. Within the UG paradigm, an unfair offer is inherently compared to a fair alternative; the more unfair an offer, the more negatively it is evaluated in terms of intentional inequity. By the same token, rejection of an unfair offer, which is costly for the responder, implies inequity aversion. A way to account for and disentangle the value and intentionality of unfair offers, is by the use of an adapted version of the UG, previously referred to as the mini-UG (Falk *et al.*, 2003). In this version, the proposer always has two alternatives by which he can divide the stakes. One of the alternatives is always an unfair division of the stake, namely eight coins for the proposer and two coins for the responder (named as 8/2 division). The intentionality manipulation is associated with the alternative division. In one condition, the proposer has a fair alternative (i.e. 5/5 division); in the second condition, the proposer has a hyperfair alternative (i.e. 2/8 division); in the third condition, the proposer has no alternative (i.e. the second alternative was also an 8/2 division). Indeed, prior behavioral research has demonstrated that responders are more willing to accept unfair offers in the no-alternative condition compared to the fair-alternative condition (Falk *et al.*, 2003; Sutter, 2007). This adapted version of the UG, where the responder is aware of the two alternatives available to the proposer, provides the possibility to compare responders' reactions to the same unfair offer (i.e. 8/2 division) under different intentionality conditions and allows the comparisons of inequity aversion (reject unfair offers independent of context) *vs* intentionality consideration (only reject unfair offers when the proposer had a better alternative).

The goal of this study was to examine neural correlates of intentionality related fairness considerations in the adapted version of the UG. We were specifically interested in behavioral and neural responses to unfair offers (8/2 division), which were offered in three context conditions (fair-, hyperfair-, no-alternative). Based on prior studies (Sutter, 2007; Güroğlu *et al.*, 2009), we predicted that acceptance of unfair offers would be higher for the no-alternative relative to the fair- and hyperfair-alternative conditions.

Based on prior studies using the classic UG, we expected involvement of the DLPFC and the insula in response to unfair offers. Along with previous findings, we expected DLPFC to elicit a regulatory function in controlling self-serving impulses to accept unfair offers and thus be more active in rejection of unfair offers (Knoch *et al.*, 2006). We expected insula involvement specifically for rejection of unfair offers (Sanfey *et al.*, 2003). The comparison of context conditions allowed us to dissociate between different alternative hypotheses regarding the role of the insula in rejecting unfair offers. If the insula have a general role in affective responses to unfair offers (i.e. inequity aversion), then they should be most active when rejecting unfair offers when the proposer has a fair alternative (Sanfey *et al.*, 2003). In contrast, if the insula are associated with error signals related to norm violations in social context, then they should be most active when rejecting unfair offers in

which the proposer has no alternative (King-Casas *et al*., 2008). The focus of this study was on responses related to unfair offers; therefore, we did not predict involvement of regions previously associated with positive affect.

The current task conditions also allowed us to examine the relative contributions of brain regions which have previously been associated with theory-of-mind and mentalizing, specifically the temporoparietal junction (TPJ) and anterior medial PFC (Decety and Lamm, 2007; van Overwalle, 2009). Rejection of an offer is costly for the responder and thus might be a more difficult decision to make than accepting an offer. Rejecting an unfair offer in the adapted UG paradigm, however, is readily justified when the alternative offer is fair or hyperfair. Whereas the classic UG elicited neural responses associated with a fast norm evaluation related to basic fairness considerations, the adapted version required mentalizing about intentions. We further reasoned that especially the condition where unfair offers are associated with no alternative options for the proposer (i.e. the no-alternative condition) would require the highest level of intentionality understanding because rejecting an unfair offer (costly decision) cannot be readily justified. Thus, we expected increased activation in TPJ and anterior medial PFC during rejection of unfair offers in the no-alternative condition.

In short, this study aimed to examine neural activity associated with responses to unfair offers within an UG paradigm that manipulates intentionality of offers. To this end, we first examined brain regions involved in context by response interactions and then examined additional brain regions that were differentially associated with a certain response (i.e. rejection or acceptance) in different contexts as well across contexts.

METHOD

Participants

Twenty-five healthy right-handed adults (15 females) between ages 18 and 25 participated in the study. Two participants were excluded from the study due to technical problems with the obtained images. The remaining 23 participants had a mean age of 20.4 years (s.d. = 1.7 years; 13 females). All participants completed a checklist confirming eligibility to take part in an MRI scan; none of them reported neurological or psychiatric impairments. The study was approved by the medical ethical committee of the Leiden University Medical Center. In accordance with their policies, all anatomical scans were reviewed by a radiologist; no anomalies were reported.

Participants completed the pen and paper version of the Raven Standard Progressive Matrices (SPM; Carpenter *et al*., 1990) to assess their inductive reasoning ability and to obtain an estimate of their intelligence quotient (IQ). The participants had above average IQ as measured by the transformed Raven SPM scores ($M = 107.39$, s.d. = 12.51). There were no gender differences in IQ [$F(1,21) < 0.01$, $P = 0.97$] and IQ was not related to behavioral performance assessed by rejection rates of unfair offers in the UG [all $r(23) < -0.37$, $P > 0.08$].

The UG task

Participants played the role of the second player in a modified version of the two-person UG, which allows to incorporate intentions behind monetary offers (Falk *et al*., 2003; Fehr *et al*., 2008). In the modified version of the UG, the proposer has to choose between a fixed set of two distributions in order to share the stake with the responder. In this study, we employed three conditions, where the unfair distribution of the proposer receiving eight coins and the responder receiving two coins (hereafter 8/2 offer) is pitted against three alternatives offers: (i) 5/5 offer (fair-alternative), (ii) 2/8 offer (hyperfair-alternative) and (iii) 8/2 offer (no-alternative).

Participants played 168 rounds of the game with anonymous age and gender matched partners. Each round was played with a new player to avoid learning and reputation effects. Only the first name and the first letter of the surname of the players were displayed on screen to ensure anonymity. Participants were told that the offers of the proposers had already been obtained in a previous part of the study. Participants were told that at the end of the session the computer would randomly select 10 rounds that would determine their total earnings. In order to emphasize the interactive character of the games with consequences for them and the other players, participants were explained that the same rule applied to proposers who would be paid later on contingent upon their decisions. At the end of the session, a screen was presented indicating the pay-off (five euros for each participant) determined by the computer. In reality, the proposals were based on behavioral pilot testing and all offers presented to the participants were computer simulated. Participants were debriefed to ensure that they believed the cover story and none of the participants expressed doubts.

Prior to the scanner task, participants practiced the task on a laptop computer for 24 trials in total. The 168 trials of the scanner task consisted of 126 trials of unfair offers (42 trials in each condition), 21 trials of fair offers and 21 trials of hyperfair offers (see Table 1). The trials were

Table 1 Number of trials per condition and offer made in each condition

Condition	Offer made	Number of trials
Fair-alternative (8/2 *vs* 5/5)	Fair (5/5)	21
	Unfair[a] (8/2)	42
Hyperfair-alternative (8/2 *vs* 2/8)	Hyperfair (2/8)	21
	Unfair[a] (8/2)	42
No-alternative (8/2 *vs* 8/2)	Unfair[a] (8/2)	42

[a]Trials that are analyzed in the neuroimaging data.

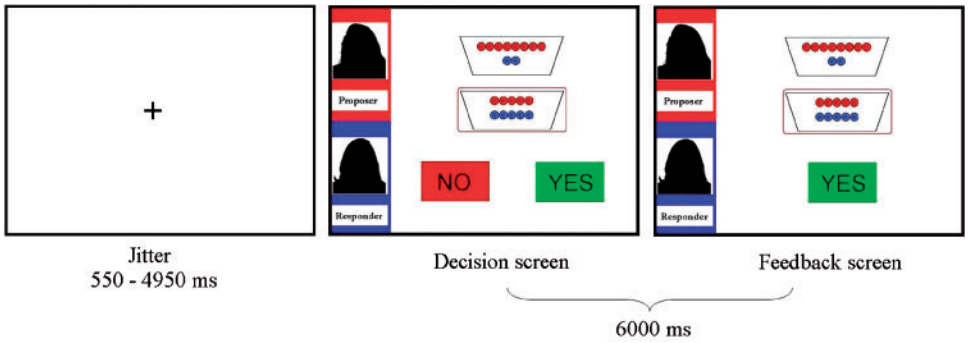

Fig. 1 Visual display and timing of the events in the scanner task in milliseconds (ms). Two offers each containing red and blue coins indicate the share for the proposer and the responder, respectively (here 8/2 *vs* 5/5). The left panel displays the name of the proposer in red (here 'proposer') and the name of the responder in blue (here 'responder'). Red encircled option indicates the offer made by the proposer (here 5/5); the responder is asked to select Yes or No to accept or reject the offer made by the proposer. The decision screen was response terminated with a maximum response time of 5000 ms and was followed by a feedback screen, which remained on the screen until 6000 ms after the start of the trial. The feedback screen presents display of given response (here 'Yes').

presented in random order in three blocks, consisting of 42 trials each and lasting ~8.3 min. Each trial lasted ~7500 ms. The trials were randomized with a jittered interstimulus interval (min = 0.55 s, max = 4.95 s, mean = 1.53 s) optimized with OptSeq2 (surfer.nmr.mgh.harvard.edu/optseq/, developed by Dale, 1999). Each trial started with a fixation after which the participants were presented with the two sets of distributions that were available to the proposer; the encircled distribution indicated the offer made by proposer (see Figure 1). Upon presentation of this screen, the responders were asked to accept or reject this offer by choosing the YES or NO button using the index or middle finger of their right hand. Participants had 5000 ms to make a decision. As soon as decision was made, feedback displaying their decision was displayed on the screen until the end of 6000 ms. In case participants failed to respond within 5000 ms, participants were presented with a screen displaying the text: 'Too late!' for the remaining 1000 ms. These occurred in <2% of the trials.

MRI data acquisition

Participants were scanned using 3.0T Philips Achieva scanner at the Leiden University Medical Center. Participants viewed the stimuli, which were projected onto a screen at the head of the scanner bore, by means of a mirror mounted on the head coil assembly. The scan sessions started with a localizer scan, followed by T2*-weighted echo-planar imaging (EPI) sequence that measures the bold-oxygen-level-dependent (BOLD) signal [TR = 2.2 s, TE = 30 ms, slice-matrix = 80 × 80, slice-thickness = 2.75 mm, slice gap = 0.28 mm gap, field of view (FOV) = 220 mm]. There were three functional runs of 200 volumes each. The first two scans were discarded to allow for equilibration of T1 saturation effects. After the functional scanning, a high-resolution T1-weighted anatomical scan and a high resolution T2-weighted matched-bandwidth high-resolution anatomical scan (same slice prescription as EPI) were obtained. Stimuli were presented and recorded using E-Prime software.

MRI data analysis

Image preprocessing and analyses were carried out using SPM5 software (http://www.fil.ion.ucl.ac.uk). Functional images were (i) slice-time corrected, (ii) realigned, (iii) spatially normalized to EPI templates and (iv) spatially smoothed using a 6 mm full-width half-maximum 3D Gaussian kernel. Movement parameters in all directions were below 1.8 mm for all participants and all scans. The data were modeled by a series of events convolved with a canonical haemodynamic response function (HRF). Each trial was modeled based on the moment of stimulus presentation (with zero duration), but our analyses only targeted the unfair offers (8/2 offers); the alternative offers were modeled separately. Each unfair offer was modeled based on the context (three levels: fair, hyperfair, or no alternative) and the participant's response (two levels: accept or reject), resulting in a 3 × 2 full factorial design. The analyses were carried out using a general linear model that included regressors for each condition. First, contrast parameter images were obtained for each individual. Consequently, these images were used in the second-level analysis of variance using the random effects model. At the group level, full factorial ANOVAs, as well as one-tailed post hoc *t*-tests, were conducted on these images.

Mean rejection levels per condition were used in regression analyses to test for brain-behavior relations. Rejection levels were not collapsed across conditions because rejection rates of unfair offers differed across conditions (see behavioral results below). The fMRI analyses did not survive whole brain corrections; they were conducted at the commonly used (Sanfey *et al.*, 2003; Tabibnia *et al.*, 2008) threshold of $P < 0.001$ uncorrected with a voxel threshold of five functional voxels, unless otherwise indicated.

Region-of-interest (ROI) analyses

Effects obtained in the full factorial ANOVAs were further examined with ROI analyses. These analyses are more powerful than whole-brain contrasts in detecting effects that are present in certain predetermined brain regions of interest, including the DLPFC, the insula, anterior medial PFC and the TPJ. ROI analyses were conducted using the Marsbar toolbox in SPM5 (Brett *et al.*, 2002; http://marsbar.sourceforce.net/).

RESULTS

Behavioral results

The dependent variable of interest in the behavioral responses was the rejection rate of unfair offers. Figure 2 displays the rejection rates of unfair offers as well as the rejection rates of alternative offers in each condition.[1] A repeated-measures ANOVA indicated significant differences between rejection rates of unfair offers across the three conditions, $F(2,44) = 44.88$, $P < 0.001$. *Post hoc* analyses showed that rejection rates in the fair- and hyperfair-alternative conditions did not differ from each other ($M = 82.96\%$, s.d. $= 23.19$ and $M = 73.04\%$, s.d. $= 26.47$, respectively; $F(1,22) = 2.67$, $P = 0.12$), whereas they were both higher than rejection rates of unfair offers in the no-alternative condition [$M = 22.97\%$, s.d. $= 29.69$; both $F(1,22) > 48.96$, $P < 0.001$]. Replicating previous findings (Sutter, 2007; Güroğlu *et al.,* 2009), these results once again show that intentions modulate acceptance of unfair offers such that the same offer is differentially rejected depending on the context in which it was made.

The rejection and acceptance of unfair offers were the primary focus of the neuroimaging analyses. However, behavioral analysis of the alternative offers showed that, as expected, almost every fair and hyperfair offer was accepted ($M = 99.92\%$, s.d. $= 1.78$ and $M = 97.90\%$, s.d. $= 3.90$, respectively). These offers were not analyzed further in the neuroimaging data.

Neuroimaging results

Context and response interaction effect. To assess the differences in the BOLD signal in the six categories of our design, we examined the interaction effect between context (three levels; hyperfair-, fair- and no-alternative) and response (two levels; accept and reject) to unfair offers with an ANOVA in the 3×2 full factorial design. In the whole brain analysis, there was a significant response × context interaction effect in bilateral insula/inferior frontal gyrus (IFG) and dorsal anterior cingulated cortex (ACC) [$F(2,117) = 7.33$, $P < 0.001$; see Table 2).

In order to further test sensitivity to the context manipulations, we conducted ROI analyses for the regions that were obtained from the interaction effect on the whole-brain analysis: bilateral insula (MNI 45, 24, −12 and −33, 21, 6) and dorsal ACC (MNI 9, 33, 24). As can be see in Figure 3, activation in the right insula was highly context dependent [$F(2,24) = 13.59$, $P < 0.001$], showing increased activation for the rejection > acceptance contrast in the no-alternative condition ($P < 0.01$) but for the acceptance > rejection contrast in the fair condition ($P < 0.01$); there was no significant difference between activation for acceptance and rejection in the hyperfair condition ($P = 0.07$). Moreover, brain activity during rejection of unfair offers in the no-alternative condition did not differ from those during acceptance of unfair offers in the fair- and hyperfair-alternative conditions. Similar patterns of interactions were found for the left insula $F(2,24) = 13.43$, $P < 0.001$ and dorsal ACC $F(2,24) = 9.52$, $P < 0.01$, and higher order comparisons revealed that these regions did not differ from each other (P's > 0.1).

These results are in favor of the hypothesis that insula activation is associated with social norm violations (or the response that is inconsistent with the context). This hypothesis is reinforced by brain behavior correlations for each condition. Negative correlations were found between the mean rejection level and BOLD activity for the reject > acceptance

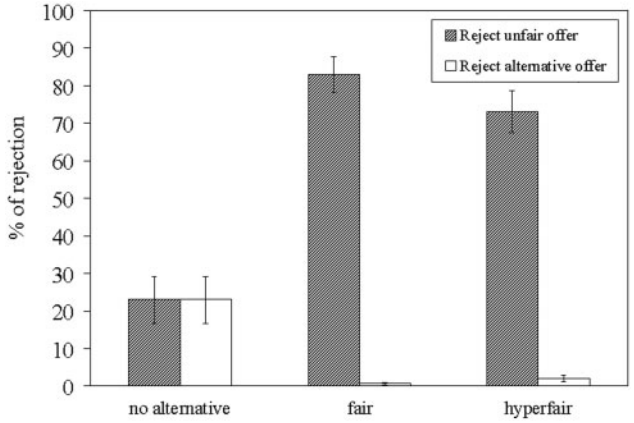

Fig. 2 Mean and standard deviations for the rejection rates of unfair and alternative offers in the three conditions.

Table 2 Areas of significant activation for the interaction between response and context in the 2×3 ANOVA design, $P < 0.001$ uncorrected, five voxel threshold

Brain region	Cluster size	Left/right	*x*	*y*	*z*	*Z*
Anterior cingulate cortex	11	R	9	33	24	3.60
Insula/IFG	43	R	45	24	−12	4.26
	9	L	−33	21	6	3.56
Posterior insula	5	R	39	−12	15	3.35
Middle temporal gyrus	6	R	60	−3	−15	3.70
Postcentral gyrus	13	R	18	−39	63	3.64
Precentral gyrus	30	L	−21	−27	60	4.24

[1] Note that in the 'no-alternative' condition the rejection rate of unfair and alternative offer are the same because they refer to the same offer.

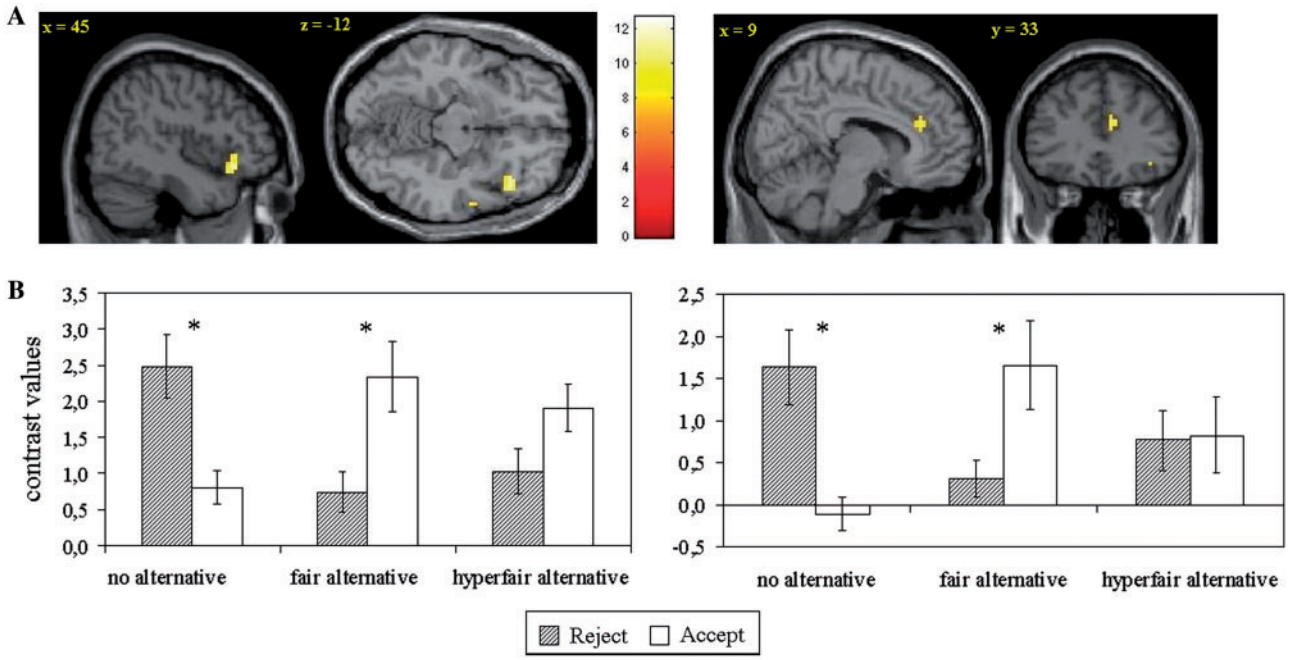

Fig. 3 (**A**) Brain regions [right insula (on left; MNI (45, 24, −12)] and dorsal ACC [on right; MNI (9, 33, 24)] of significant interaction in the 2 × 3 full factorial design and (**B**) contrast values in these regions (right insula on left; dorsal ACC on right) for acceptance and rejection of unfair offers in the three conditions. Significant differences between brain activity for acceptance and rejection of unfair offers in each condition are indicated with an asterisk. (Results for left insula are not shown).

contrast in left insula (no-alternative $r = -0.56$, $P < 0.05$, fair-alternative $r = -0.56$, $P < 0.05$, and hyperfair-alternative $r = -0.48$, $P < 0.05$), right insula (no-alternative $r = -0.67$, $P < 0.01$, fair-alternative $r = -0.66$, $P < 0.01$, and hyperfair-alternative $r = -0.52$, $P < 0.05$) and dorsal ACC (no alternative $r = -0.64$, $P < 0.01$). Thus, brain activity in these regions was higher when participants responded in a way that that they do not usually do, that is, when they violated their own norms for behavior.

Context effect in rejection and acceptance. The next question concerned the differences in neural activation that was specific for a certain response, by comparing rejection and acceptance trials for the three context conditions separately. These analyses were performed to detect differential activation patterns in regions that were not detected in the full factorial ANOVA.

For unfair rejection trials, a whole-brain analysis of variance examining effects of context (three levels: fair-, hyperfair- and no-alternative) yielded a main effect of context in the medial PFC/ACC and right TPJ, $F(2,61) = 7.75$ (see Table 3). In these brain regions we conducted post hoc ROI analyses to better understand the brain activity associated with rejection in relation to acceptance trials. Results showed that in right TPJ and the medial PFC/ACC brain activity related to rejection of unfair offers in the no-alternative condition was associated with higher activity than rejection of unfair offers in the fair- and hyperfair-alternative conditions [$F(2,34) = 13.94$, $P < 0.001$ and $F(2,36) = 20.93$, $P < 0.001$, respectively; see Figures 4A and 4B].

Table 3 Areas of significant activation for the rejection of unfair offers in the no alternative condition *vs* fair and hyperfair conditions in the 1 × 3 ANOVA design, $P < 0.001$ uncorrected, five voxel threshold

Brain region	Cluster size	Left/right	*x*	*y*	*z*	*Z*
Anterior cingulate cortex	18		−3	42	12	4.24
	24		0	30	3	3.95
Insula	68	R	54	21	−6	4.02
		R	48	24	−12	3.77
		R	42	36	−9	3.50
Temporoparietal junction	43	R	63	−48	33	4.09

Thus, contrary to the insula, activation in right TPJ and medial PFC/ACC was specific for the no-alternative rejection trials. Indeed, these are the conditions that are associated with the highest mentalizing demands in terms of violating expectations of others. As expected, ROI activity in these regions was again correlated with individual performance: higher levels of activity in right TPJ as well as in medial PFC/ACC were related to lower rejection levels, specifically for the no-alternative condition (TPJ no-alternative $r = -0.50$, $P < 0.05$; medial PFC/ACC no-alternative $r = -0.52$, $P < 0.05$).

For acceptance trials, the ANOVA resulted in a significant main effect of context in the precuneus and the posterior and anterior insula [$F(2,56) = 7.83$]. Thus, no additional regions of interest were identified for the acceptance of unfair offers beyond those obtained in the 3 × 2 interaction described above.

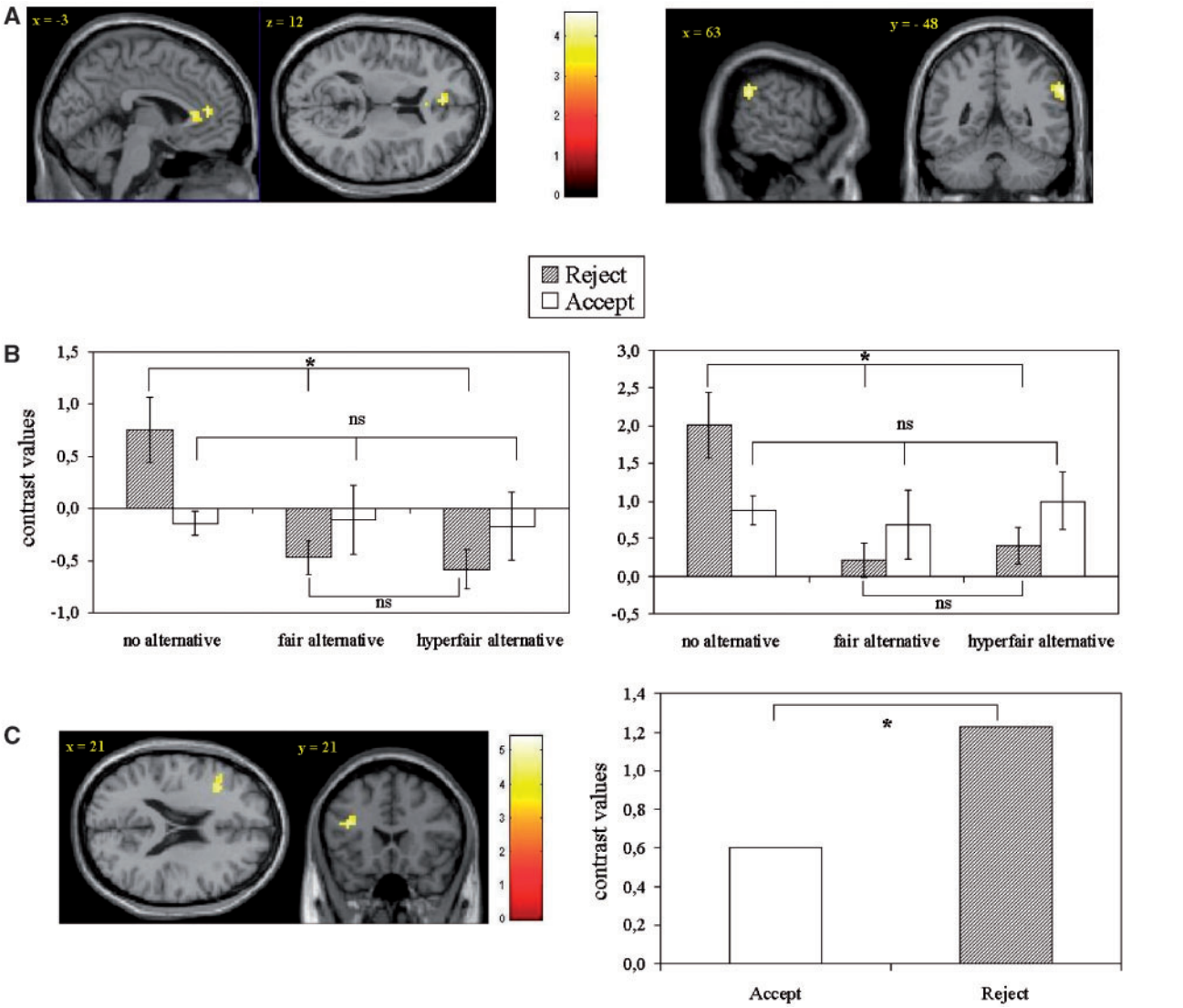

Fig. 4 (**A**) Brain regions [ACC (on left; MNI −3, 42, 12) and TPJ (on right; MNI 63, −48, 33)] of context effect during rejection of unfair offers and (**B**) contrast values and standard errors in these regions (ACC on left; right TPJ on right) for acceptance and rejection of unfair offers in the three conditions. (**C**) Brain region (left DLPFC; MNI −36, 21, 21) and contrast values and standard errors for rejection greater than acceptance of unfair offers in the three conditions. Significant differences between brain activity are indicated with an asterisk; nonsignificant differences are indicated with ns ($P > 0.05$).

Response main effect. Finally, we were interested to see if additional brain regions could be identified for response related brain activity across the three contexts. For this purpose, main effect of response to an unfair offer was examined by the rejection *vs* acceptance contrast where the three context conditions were collapsed. The contrast of rejection > acceptance for unfair offers yielded activity in left DLPFC (MNI − 36, 21, 21), $T(22) = 3.50$ (see Table 4). ROI analyses were conducted to further examine the main effect of rejection in left DLPFC (see Figure 4C). Repeated measures ANOVAs with response (two levels: acceptance *vs* rejection) and context (three levels: no-, fair- and hyperfair-alternative) as within subjects factors yielded a main effect of response [$F(1,11) = 7.35$, $P < 0.05$], but no

Table 4 Areas of significant activation for the acceptance versus rejection contrast, $P < 0.001$ uncorrected, 5 voxel threshold

Brain region	Cluster size	Left/right	*x*	*y*	*z*	*Z*
Rejection > acceptance						
Dorsolateral prefrontal cortex	26	L	−36	21	21	3.88
Insula	6	L	−27	24	−6	3.54
Postcentral gyrus	6	L	−36	−33	48	3.69
Precentral gyrus	15	L	−42	0	42	3.65
SMA	9	L	−9	12	45	3.54
Supra marginal gyrus	31	R	54	−24	39	4.27
	45	L	−54	−24	51	3.80
Acceptance > rejection						
Anterior cingulate cortex	40		0	36	3	3.59
Middle cingulate cortex	7		0	−30	42	3.75
Supra marginal gyrus	63	R	63	−39	42	4.03
Middle frontal gyrus	34	R	27	33	42	4.00

main effect of context [$F(2,22) = 1.54$, $P = 0.24$], and no interaction effect of response and context [$F(2,22) = 2.82$, $P = 0.09$ with Huynh–Feldt correction]. This finding suggests that rejection of an unfair offer is paired with regions previously associated with regulation and control. If self-gain is the motivating factor for an underlying tendency to accept all offers (because that is the only way players will receive gains), rejection of an unfair offer possibly involves control of this tendency. A comprehensive list of brain areas involved in the contrast of rejection > acceptance and acceptance > rejection are depicted in Table 4.

DISCUSSION

This study set out to test the role of intentionality consideration in fairness judgments. Consistent with prior studies, rejection of unfair offers was dependent on the intentions behind unfair offers (Sutter, 2007; Güroğlu *et al.*, 2009). That is, participants more often rejected unfair offers when the proposer could also have selected a fair- or hyperfair-alternative. These findings demonstrate that participants valued fairness even when this occurred at the cost of their own benefit (Fehr and Schmidt, 1999; Falk *et al.*, 2003). In contrast, when proposers had no alternative but to offer an unfair division, rejection rates were significantly reduced. It should be noted that rejection rates were still around 30% even in the no-alternative condition, which can be explained by individuals' tendency towards inequity aversion (Fehr and Schmidt, 1999). Alternatively, it is possible that no-alternative unfair offers potentially still involve bad intentions (which were hidden by the no-alternative character of the task). Rejection of unfair offers, however, decreased from 75 to 30% when proposers had no other option than to offer an unfair division, showing that intentions behind unfair offers modulate rejection (Falk *et al.*, 2003; Sutter, 2007). These task manipulations allowed us to investigate the neural correlates of inequity aversion *vs* intentionality understanding associated with different decisions (i.e. accept *vs* reject).

Social norms

Consistent with prior studies, the insula, as well as the dorsal ACC, was engaged in fairness judgments (Sanfey *et al.*, 2003). The direction of activation, however, was different from prior studies and appeared dependent on task context and associated with behavioral outcomes. Higher levels of insula activity were obtained during rejection of unfair offers in the no-alternative condition, as well as during acceptance of unfair offers when there was a fair- or hyperfair-alternative. Furthermore, insula activation was associated with performance such that those individuals with lower rejection rates of unfair offers had higher insula activity when they rejected these offers. In other words, levels of brain activity were higher when participants engaged in behavior that they usually do not show, confirming the role of the insula in norm violations of behavior (Montague and Lohrenz, 2007).

This interpretation is supported by a study examining the neural basis of social norm compliance (Spitzer *et al.*, 2007). In this study, participants played two versions of the UG as proposers: a control UG, where the responder could not reject the proposer's offer, and a punishment UG, where the responder could punish the proposer by taking back coins. Insula activity difference in the punishment and control conditions was found to correlate with differences in transferred amounts in the two games, as well as with Machiavellianism scores of individuals, where high Machiavellianism indicates strong deviations from the norm. Furthermore, diminished activity of the anterior insula was found in borderline disorder patients with pathological disturbance of norms related to social gestures (King-Casas *et al.*, 2008). Unfair offers might be perceived as violations of social norms; insula activity has indeed been associated with receiving unfair offers (Sanfey *et al.*, 2003). Rejection of unfair offers might further signal an internal error related to the conflict between the self-gain maximization goal (i.e. accepting the offer) and the negative emotional response (i.e. rejecting the offer). When intentionality related information is incorporated into fairness judgments, norms become context-dependent. In other words, external information related to context is crucial in defining norms and norm-violations. In the no-alternative condition, accepting an unfair offer might become the norm because the proposer has no other alternative, whereas in the fair-alternative condition, rejection remains as the norm. It seems unlikely that these findings are simply the result of probability effects, because the hyperfair-alternative condition did not show this differentiation, whereas behavior was comparable to the fair-alternative condition. It should be noted that, unlike in the fair condition, neural activity related to acceptance and rejection of unfair offers does not differ in the hyperfair condition. It is possible that rejection of an unfair offer in the hyperfair condition is not as readily justifiable as in the fair condition because one can understand that the proposer does not want to make a 2/8 offer. In this sense, the hyperfair condition remains relatively ambiguous compared to the fair condition and the interpretation of the results remains challenging.

The insula activation in the current study shows a different pattern compared to results previously reported by Sanfey and colleagues (2003) using a classic UG. These differences could be related to design changes. First, in the original UG there is a variation of unfair offers (i.e. 9/1, 8/2, 7/3 and 6/4) which might render the 8/2 offer to be perceived differently than the 8/2 offer in the design employed here. In the current design, all unfair offers refer to the same offer (i.e. 8/2 distribution) where context of the offer shapes the way it is perceived. Second, the insula activation in the Sanfey *et al.* (2003) study was related to the presentation of the unfair offer, but not to the actual

behavioral response, which occurred at least 6 s later. In our design, participants were presented with alternatives as well as the offer at the same time point and were asked to give a behavioral response upon presentation of the offer. Our findings suggest that norm violations should be interpreted in context and should be related to typical behavior (e.g. acceptance of unintended unfair offer as norm). In the current study, evaluation of an offer and decision-making were intertwined processes. Future experiments should aim to dissociate these processes and better understand the role of insula in perception of norm violations (e.g. receiving an unfair offer) and evaluation of unfairness in context (e.g. including information on intentionality). Conducting studies using other modalities such as event related potential (ERP) measurements and source localization can prove fruitful in resolving these issues and obtaining a better idea of the temporal patterns in brain activity in decision-making.

Insula activity related to rejection of unfair offers in the no-alternative condition and acceptance of offers in the fair- and hyperfair-alternative conditions was also accompanied by dorsal ACC activity. Studies on the specialization of ACC in social cognition and decision-making point out that dorsal ACC is involved in response conflict and error monitoring (van Overwalle, 2009). The involvement of the ACC in unfair offers and their rejection is possibly due to its role in monitoring and detection of conflicts between the emotional and cognitive components of social decision-making (Sanfey *et al*., 2003). Furthermore, the anterior insula (AI)/ACC brain network has been shown to be activated in pain-related empathy for fair-acting (but not for unfair-acting) others, yielding further support for the AI/ACC network involvement in norm violations (Singer *et al*., 2006).

Mentalizing

An additional goal of the current study was to examine the role of mentalizing-related brain regions in social decision-making involving intentionality considerations. Comparison of brain activity related to rejection of unfair offers across context with differing intentionality in received offers yielded the involvement of the medial PFC/ACC and rTPJ. Activity in these regions was higher during rejection of unfair offers when the proposer had no-alternative compared to when the alternative was a fair or hyperfair distribution of the stake. In previous research, the mPFC has been associated with various aspects of social cognition, such as mentalizing, self-perception, as well as action and outcome monitoring (Amodio and Frith, 2006). Within this region, the mPFC/ACC was found to be important for complex emotional and social behavior involved in, for example, social interactions (Rudebeck *et al*., 2008). Particularly ventral ACC is shown to be involved in theory of mind beliefs, self-referential thinking and emotionality (van Overwalle, 2009). Similarly, rTPJ plays an important role in higher level social cognitive processing, particularly in terms of its involvement in understanding others' mental states and feelings of empathy (Saxe and Wexler, 2005; Decety and Lamm, 2007). Rejecting an unfair offer in the fair- and hyperfair-alternative conditions might be taken for granted, whereas rejection of an unfair offer in the no-alternative condition cannot be as readily justified. In this sense, rejection of an unfair offer in the no-alternative condition involves higher levels of mentalizing and attribution of mental states. Furthermore, negative correlations between brain activity in the mPFC and TPJ regions and behavioral performance suggest that those participants who often accepted unfair offers might engage in higher levels of mentalizing during rejection of these offers, possibly due to higher levels of consideration for violating others' expectations. Both of these brain regions have been implicated to play a role in moral judgments, with particular role for TPJ in belief formation related to intentionality and for mPFC in response conflict (Young *et al*., 2007; van Overwalle, 2009). Thus, the findings might be related to feelings of guilt associated with the conflict of rejecting an unfair offer (i.e. punishing the proposer), even though the proposer had no alternative (i.e. was helpless).

DLPFC: control of rejection

Our findings revealed DLPFC involvement with rejection of unfair offers. Notably, this pattern of activation was independent of context, which suggests a role of regulation and control in rejecting unfair offers in general. Goals of self-gain would require participants to accept every unfair offer, which might require executive control to suppress. In this sense, our findings are consistent with results reported by Knoch and colleagues (2006) regarding the role of DLPFC in overriding self-interest motives during rejection of unfair offers. Once again, it is crucial for future research to untangle neural activity related to perception of unfair offers and those related to behavioral responses to unfair offers. It is also necessary for future research to examine emotional reactions to perception of unfairness as well as motivations behind behavioral responses to unfairness.

CONCLUSION

Taken together, the results of this study point out a role for several brain regions in social decision-making related to fairness considerations. We have focused on intentionality considerations as the significant context characteristic in fairness considerations. Extending previous research findings, we have demonstrated that brain networks involved in fairness considerations are highly context-dependent. Our findings show that decision-making is modulated by context such that the same behavioral response (e.g. rejection of an unfair offer) is related to differential neural activity patterns depending on intentionality of offers.

The insula and dorsal ACC appeared most sensitive to context manipulations by showing increased activation when participants act against socially accepted norms. Our

findings therefore support the hypothesis that these regions are sensitive to social norm violation, and disconfirm the hypothesis that these regions are simply responsive to inequity aversion. In addition, this was the first study to demonstrate the involvement of the medial PFC and TPJ in the UG, by manipulating intentionality and thereby putting higher demands on mentalizing.

Future research needs to focus on disentangling to what extent these brain regions involved in social decision-making are related to affective *vs* cognitive or automatic *vs* deliberative systems involved in decision-making (Sanfey and Chang, 2008). Studying these dual systems also present challenges to the study of neurobiological underpinnings of decision-making. Possibly, this direction can benefit from combining EEG and neuroimaging studies (Frank *et al.*, 2009).

REFERENCES

Amodio, D.M., Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Review Neuroscience*, *7*, 268.

Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.

Carpenter, P.A., Just, M.A., Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404–431.

Decety, J., Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *Neuroscientist*, *13*, 580–93.

Falk, A., Fehr, E., Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, *41*, 20–26.

Fehr, E., Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*, 785.

Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*, 817–68.

Frank, M.J., Cohen, M.X., Sanfey, A.G. (2009). Multiple systems in decision making: a neurocomputational perspective. *Current Directions in Psychological Science*, *18*, 73–7.

Güroğlu, B., van den Bos, W., Crone, E.A. (2009). Fairness considerations: increasing understanding of intentionality in adolescence. *Journal of Experimental Child Psychology*, *104*, 398–409.

Güth, W., Schmittberger, R., Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*, 367.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., Montague, P.R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, *321*, 806–10.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*, 829–832.

Lee, D. (2008). Game theory and neural basis of social decision making. *Nature Neuroscience*, *11*, 404.

Montague, P.R., Lohrenz, T. (2007). To detect and correct: norm violations and their enforcement. *Neuron*, *56*, 14.

Phan, K.L., Wager, T.D., Taylor, S.F., Liberzon, I. (2004). Functional neuroimaging studies of human emotions. *CNS Spectrums*, *9*, 258–66.

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

Rilling, J.K., King-Casas, B., Sanfey, A.G. (2008). The neurobiology of social decision-making. *Current Opinion in Neurobiology*, *18*, 159.

Rudebeck, P.H., Bannerman, D.M., Rushworth, M.F.S. (2008). The contribution of distinct subregions of the ventromedial frontal cortex to emotion, social behavior, and decision making. *Cognitive, Affective & Behavioral Neuroscience*, *8*, 485–97.

Sanfey, A.G., Chang, L.J. (2008). Multiple systems in decision making. *Annals of the New York Academy of Sciences*, *1128*, 53–62.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*, 1755–8.

Saxe, R., Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, *43*, 1391.

Singer, T., Critchley, H.D., Preuschoff, K. (2009). A common role of insula in feelings, empathy, and uncertainty. *Trends in Cognitive Sciences*, *13*, 334–40.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., Frith, C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*, 466.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., Fehr, E. (2007). The neural signature of social norm compliance. *Neuron*, *56*, 185.

Sutter, M. (2007). Outcomes versus intentions: on the nature of fair behavior and its development with age. *Journal of Economic Psychology*, *28*, 69.

Tabibnia, G., Satpute, A.B., Lieberman, M.D. (2008). The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, *19*, 339–47.

van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, *30*, 829–58.

Wout, M. van't, Kahn, R.S., Sanfey, A.G., Aleman, A. (2005). Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. *NeuroReport*, *16*, 1849–52.

Wout, M. van't., Kahn, R.S., Sanfey, A.G., Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research*, *169*, 564–8.

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