Disentangling self- and fairness-related neural mechanisms involved in the Ultimatum Game: an fMRI study.

Corrado Corradi-Dell’Acqua**†, Claudia Civai**†, Raffaella I. Rumiati† and Gereon R. Fink§¶

† Cognitive Neuroscience Sector, Scuola Internazionale Superiore di Studi Avanzati (SISSA/ISAS), Trieste, Italy.
‡ Swiss Centre for Affective Sciences, University of Geneva, Switzerland
§ Cognitive Neurology Section, Institute of Neuroscience and Medicine (INM-3), Research Center Juelich, Juelich, Germany.
¶ Department of Neurology, University Hospital Cologne, Cologne University, Cologne, Germany.
* These authors contributed equally to the study.

Running Head: Self versus fairness mechanisms in the ultimatum game

Correspondence concerning this article should be addressed to:
Corrado Corradi-Dell’Acqua, PhD, Claudia Civai
Cognitive Neurology Section, Institute of Neuroscience and Medicine (INM-3), Research Center Juelich, Leo-Brandt Strasse, D-52428, Juelich, Germany.
E-mail: Corrado.Corradi@unige.ch; civai@sissa.it

© The Author (2012). Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com
Manuscript information:

Abstract word count: 198

Manuscript word count (excluding references, tables and figure legends): 4872

Number of Figures: 4 (color figures 2)

Number of Tables: 1

Supplementary Materials: one text file.

Acknowledgments. We are grateful to all our colleagues at the Institute of Neuroscience and Medicine, Research Center Jülich. In particular, we would like to thank Edna Cieslik, Simone Vossel, Qi Chen, Ralph Weidner and Shahram Mirzazade for being so collaborative proposers. Finally, we would like thank all our volunteers for their participation in the study.
Abstract

Rejections of unfair offers in the Ultimatum Game are commonly assumed to reflect negative emotional arousal mediated by the anterior insula and medial prefrontal cortex (Sanfey et al., 2003; Koenigs & Tranel, 2007). We aimed to disentangle those neural mechanisms associated with direct personal involvement (“I have been treated unfairly”) from those associated with fairness considerations, such as the wish to discourage unfair behavior or social norm violations (“this person has been treated unfairly”). For this purpose, we used fMRI and asked participants to play the Ultimatum Game (UG) as responders either for themselves (Myself) or on behalf of another person (third-party, Civai et al., 2010). Unfair offers were equally often rejected in both conditions. Neuroimaging data revealed a dissociation between the medial prefrontal cortex, specifically associated with rejections in the Myself condition, thus confirming its role in self-related emotional responses, and the left anterior insula, associated with rejections in both Myself and Third-Party conditions, suggesting a role in promoting fair behavior also towards third-parties. Our data extend the current understanding of the neural substrate of social decision making, by disentangling the structures sensitive to direct emotional involvement of the self from those implicated in pure fairness considerations.
Disentangling self- and fairness-related neural mechanisms involved in the Ultimatum Game: an fMRI study.

In the last decades, studies in the field of economics reported systematic violations of classical economic theories’ predictions, which see maximization of one’s monetary gain as the driving principle of decision making (Von Neumann and Morgenstern, 1947). One example is the Ultimatum Game (UG), in which one player (the proposer) makes an offer to a second player (the responder) on how to divide an amount of money; the responder can either accept (i.e. the money is divided as suggested), or reject (i.e., both players get no money) the offer. Classical economic theory posits that the responder should accept every offer (“few is better than nothing”) and that the proposer, consequently, should offer the smallest amount of money possible. However, behavioral findings describe the responder likely to reject offers considered unfair and the proposer more prone to divide the money equally.

Pillutla and Murnighan (1996) suggested that negative emotions (e.g., anger, frustration) underlie responder’s behavior: in particular, the unfair treatment evokes a negative emotional reaction which, in turn, leads to rejections (wounded pride/spite model). Evidence supporting this model arise by van’t Wout et al. (2006), who measured skin conductance response (SCR) as an index of emotional arousal (Boucsein, 1992), and found increased SCR when responders were about to reject (as opposed to accept) unfair UG offers. Furthermore, Harlé and Sanfey (2007), affected responders’ emotional status prior to the game through the presentation of emotionally-salient video-clips and found that rejections increased following the presentation of sad (but neither happy nor neutral) movies. Finally, Crockett et al. (2008) reported increased rejections in those participants who, following acute tryptophan depletion, presented low levels of serotonin, a neurotransmitter involved in impulse regulation.
The UG is, for its own definition, a self-centered task, in which the person accepting/rejecting the proposers’ division is also the direct target of an unfair treatment. Thus, in all the studies reviewed above, the unfairness correlates with the amount of anger/frustration triggered in the responder, leaving open the issue of whether rejections: (a) are reactions to a self-directed unfair treatment (“I have been treated unfairly”) which, consistently with the wounded pride/spite model, evokes increased anger and frustration; or (b) are driven by pure considerations about fairness (“this person has been treated unfairly”), that is by the integration of those cognitive, emotional and motivational mechanisms which lead to the discouragement of social norm violations (Moll et al., 2008). Civai et al. (2010) recently attempted to disentangle self- and fairness-related effects by asking participants to play as responders in a modified version of the UG in which the unfair bargaining was directed not to them personally (as in the classical UG), but to an unknown person. Since in this “Third-Party” UG the responder was not the victim of an unfair treatment, the effect of anger/frustration in the choice was hypothesized to be diminished. Still, the offers in the Third-Party UG were as unfair as those in the classical (“Myself”) UG and the responder could, according to the game’s rules, accept/reject them. The analysis of SCR and of emotional ratings confirmed stronger negative emotional arousal in the Myself than in the Third-Party UG, especially during the rejections; however, the amount of rejections was significantly modulated by the unfairness of the offer, and not by the target of the offer. The data from Civai et al. (2010) suggest that rejections are predominantly driven by fairness sensitivity and that the strong negative emotional reaction seems to be elicited exclusively by the self-directed unfairness.

It is still unclear how self- and fairness-related effects in UG relate to the brain. Investigations on the classical UG implicate the anterior portion of the insular (AI) and cingulate (ACC) cortex, and the dorsolateral (DLPFC) and medial (MPFC) aspects of the prefrontal cortex (Sanfey et al., 2003; van’t Wout et al., 2005; Knoch et al., 2006, 2008; Koenigs and Tranel, 2007; Tabibnia et al., 2008;
Moretti et al., 2009; Güroğlu et al., 2010, 2011; Baumgartner et al., 2011). However, the exact role played by this network in the responder’s reaction is still under debate. For instance, AI and ACC have been associated with negative emotions such as disgust, anger, fear and pain (Damasio et al., 2000; Calder et al., 2001; Wicker et al., 2003; Corradi-Dell'Acqua et al., 2011), as well as with monitoring one’s physiological responses to affective events (SCR and hearth beat – Critchley et al., 2000, 2004; Patterson et al., 2002). Thus, the involvement of these regions in rejections might be reflective of the anger/frustration elicited by self-directed unfairness (Sanfey et al., 2003). On the other hand, recent accounts suggest that AI and ACC might mediate the integration of emotional, cognitive and motivational processes (Craig, 2009; Singer et al., 2009; Lamm and Singer, 2010) and play a critical role in detecting and reacting to social norm violations (Spitzer et al., 2007; Rilling et al., 2008; King-Casas et al., 2008; Strobel et al., 2011). It is therefore plausible that the rejection-related activity in these regions reflects the wish to sanction unfairness irrespective of the person to which it is addressed. As for DLPFC and MPFC, studies testing classical UG concur in interpreting the involvement of these regions in terms of executive control, goal-maintenance, and the monitoring/control of one’s emotional responses (van’t Wout et al., 2005; Knoch et al., 2006, 2008; Koenigs and Tranel, 2007; Moretti et al., 2009; Güroğlu et al., 2010; Baumgartner et al., 2011). These interpretations leave open the possibility of prefrontal regions monitoring/controlling those emotional responses elicited by self-related unfair treatment (see Koenigs and Tranel, 2007, for MPFC) but also promoting culture-dependent fairness goals in monetary bargaining (see Knoch et al., 2006, 2008; Baumgartner et al., 2011, for DLPFC).

We used fMRI and engaged healthy participants in the paradigm described by Civai et al. (2010). Subjects performed either the UG or a control task (Free-Win [FW]), in which they accepted/rejected money provided by the computer. Both UG and FW tasks comprehended offers addressed to either oneself or a Third-Party. FW shares many properties with the UG (e.g.,
self/other-reflection, receipt of monetary value, etc.), except the fact that the money received is the result of an unfair treatment. Furthermore, within UG, we distinguished between trials which were accepted/rejected by the participants (participants seldom reject FW offers; see behavioral results and Civai et al., 2010). This constitutes a 3 x 2 design with TASK (UG Rejections, UG Acceptances, FW) and TARGET (Myself, Third-Party) as factors and 6 conditions: URm, rejected trials when playing UG for oneself; UAm, accepted trials when playing UG for oneself; FWm, FW task addressing oneself; and, respectively, Urt, UAt, FWt, Third-Party versions of UG/FW. Of crucial interest are the functional properties of regions previously associated with the classical UG (e.g., AI, ACC, MPFC, DLPFC). If these regions code those negative emotional reactions due to a direct exposure to an unfair treatment, they should be significantly associated with the TASK*TARGET interaction, as increases of neural activity for UG (relative to FW) should be observed for the Myself but not for the Third-Party. Alternatively, if the neural activity of these regions relates exclusively to fairness, then their involvement in the UG should not be specific for the Myself, but should be observed also for the Third-Party.
Materials and Methods

Participants
Twenty-three (9 females, age: 18-35 years, average = 23.5) subjects took part in the experiment. None of the participants had any history of neurological or psychiatric illness. Written informed consent was obtained from all subjects, who were naïve to the purpose of the experiment. The study was approved by the local ethics committee.

Task and Stimuli
Task, stimuli and experimental set-up were similar to the ones employed in Civai et al. (2010). Participants underwent one session of thirty minutes. The experimental instructions (see Supplementary Materials for an English-translated instruction sheet) can be subsumed as follows: another participant (i.e., the proposer) was given a 10 € note at each trial, and he/she had to split this money with him (responder). In the Myself condition, if participants accepted the offer, the money would be divided as suggested by the proposer whereas if they rejected the offer, none of the players would get any of the money. In the Third-Party condition, if participants accepted the offer, the money would be divided (as suggested by the proposer) between the individuals acting as proposer and responder in the next experimental session; if they rejected the offer, these individuals would get no money at all; in either case, neither the proposer nor the participant, would get any money related to this trial (Figure 1a).

[Insert Figure 1 about here]
Although participants were told that they were interacting with a human proposer, they were presented with offers defined \textit{a priori} by the experimenter. These could be 1, 2, 3, 4 or 5 € out of 10 (in “1 € out of 10” the responder is offered only 10% of the money at stake). UG trials were intermingled by trials of a control (Free-Win [FW]) task, in which they were offered the same amount of money as in the UG (1, 2, 3, 4, or 5 €); however, this was not a partition between two players. In the Myself condition, participants could accept the FW offer and keep the money, or reject it and get no money. In the Third-Party condition, participants could accept the FW offer and the individual acting as responder in the upcoming experimental session would receive the money; if participants rejected the offer, the next responder received no money. In either case participants received no money related to this trial (see Figure 1a).

In order to strengthen the participants’ belief that they were facing a human fellow, they were introduced prior to the experimental session to a collaborator of the experimenter who pretended to act as the proposer. Furthermore, participants were told that the proposer would receive feedback only at the end of the experiment (i.e., “covered” UG, which prevents strategic use of rejections – Zamir, 2001; Civai et al., 2010). Participants were informed that their compensation for participating in the experiment would be proportional to the amount of money gained during the Myself condition. Moreover, they knew that a proportion of the money gained on behalf of third parties would be given to the next players; they were also informed that, following the same principle, their starting stakes were proportions of the money that previous players had split on their behalf. Irrespective of task performance, participants received the same amount of money as compensation after completion of the experiment. Finally, after the whole experimental session an informal debriefing was carried out to assess whether participants believed whether offers were genuinely human. None of the participants exhibited doubts regarding the cover story.
Experimental Set-Up

Participants lay supine in the MR scanner with their head fixated by firm foam pads. Stimuli were presented using Presentation 11.0 (Neurobehavioral Systems) and projected to a VisuaStim Goggles system (Resonance Technology). Behavioral responses were recorded by pressing the corresponding keys of an MRI-compatible response device (Lumitouch, Lightwave Medical Industries, CST Coldswitch Technologies).

For each experimental trial, participants were first presented with the offer (“I offer you/the next participant 2 € out of 10”) for 4500 msec, followed by a blank screen ranging from 4750 msec to 6750 msec with an incremental step of 500 msec. The question “Do you accept?” was then presented for 2000 msec, by which time the participant had to give a response by button press. Trials were followed by an inter-trial interval ranging from 4750 msec to 6750 msec with an incremental step of 500 msec (see Figure 1b). Each experimental session comprised 105 randomized trials, including 100 experimental trials [2 (Ultimatum Game, Free Win) * 2 (Myself, Third-Party) * 5 (1, 2, 3, 4, 5 €) * 5 repetitions] and 5 “null events” in which an empty screen replaced the stimuli.

fMRI data acquisition

A Siemens Trio 3-T whole-body scanner was used to acquire both T1-weighted anatomical images and gradient-echo planar T2*-weighted MRI images with blood oxygenation level dependent (BOLD) contrast. The scanning sequence was a trajectory-based reconstruction sequence with a repetition time (TR) of 2200 msec, an echo time (TE) of 30 msec, a flip angle of 90 degrees, a slice thickness of 3 mm, and no gap between slices. For each subject, 878 volumes were acquired during the whole experimental session.
**Imaging processing**

Image processing and statistical analysis were performed using the SPM8 software package (http://www.fil.ion.ucl.ac.uk/spm/). For each subject, the first six volumes were discarded. To correct for head motion, the functional images were then realigned to the new first functional image (Ashburner and Friston, 2004), normalized to a template based on 152 brains from the Montreal Neurological Institute (MNI) at a 2x2x2 mm voxel size, and then smoothed by convolution with a 8 mm FWHM Gaussian kernel.

Data were then fed into a first level analysis using the general linear model framework (Kiebel and Holmes, 2004) implemented in SPM8. On the first level, for each individual subject, we fitted a linear regression model to the data. For the UG only, we distinguished between rejected and accepted offers. This yielded a 3 x 2 factorial design with 6 conditions. For each of these conditions we modeled independently the onset of the offer and the onset of the text string prompting a button press (see Figure 1b) through a stick function. For each of the resulting 12 vectors, we accounted for putative linear changes of neural activity across all repetitions by using the time modulation option implemented in SPM, which creates a new regressor in which the trial order is modulated parametrically. Furthermore, regressors testing the parametric modulation of the factor GAIN were included: distinct regressors were modeled for the two onsets within the trial structure (offer, response – see Figure 1b), the two levels of TARGET (Myself, Third-Party) and for TASK which was UG and FW, but not for different responses within UG trials. This yielded 32 vectors [12 stick functions + 12 time modulation vectors + 8 gain modulation vectors], each of which was convolved with a canonical haemodynamic response function and associated with a vector describing its first order time derivative. Finally, we included six differential realignment parameters as regressors. Low-frequency signal drifts were filtered using a cutoff period of 128
sec. Critically, response regressors (e.g., $URm$) correlate strongly with regressors testing for response-independent effects of GAIN (see behavioral results). By modeling both of them, we ruled out potential confounding effects of the correlated regressor, and insured that our results (if any) could be uniquely interpreted (Andrade et al., 1999).

On the second level, we focused on those parameter estimates from the first level associated with the 6 conditions of our 3 x 2 design, exclusively when the offer was presented. These images were then fed into a flexible factorial design with a within-subject factor of six levels using a random effects analysis (Penny and Holmes, 2004).
Results

Behavioral results

For each subject and for each condition, the rejection rates were calculated across all 5 repetitions, and used in a 2 TASK (UG, FW) x 2 TARGET (Myself, Third-Party) x 5 GAIN (1-5 €) Repeated Measures ANOVA. Results indicated a significant main effect of TASK ($F(1, 22) = 123.89, p < 0.001$), with the UG leading to a larger number of rejections than the FW, as well as a main effect of GAIN ($F(4,88) = 58.73, p < 0.001$), with lower offers being rejected more often than higher offers. These effects were, however, driven by a TASK*GAIN interaction, which was also significant ($F(4,88) = 63.44, p < 0.001$), suggesting that lower offers were rejected significantly more often than higher offers in the UG but not in the FW (see Figure 2). None of the remaining effects of the ANOVA was significant. This statistical analysis was performed using SPSS 11.5 Software (SPSS Inc., Chertsey UK).

[Insert Figure 2 about here]

Neural Activations

Table 1 reports areas of activations which exceeded a height threshold of $t > 3.17$ (corresponding to $p < 0.001$, uncorrected). With this height threshold, in our dataset, clusters associated with a $p < 0.05$ corrected for multiple comparisons across the whole brain were observed with an extent threshold > 176 voxels (Friston et al., 1993). Furthermore, we focused our analysis on those structures previously implicated in rejection effects in the UG: AI, ACC, MPFC and DLPFC. We created a small volume of interest including those regions in both hemispheres which, according to the AAL atlas (Tzourio-Mazoyer et al., 2002) corresponded to insula, ACC and superior/middle...
frontal gyri in both their lateral and medial aspects (regions F1, F1M, F1MO, F2). In this small volume clusters associated with $p < 0.05$, corrected, will be $> 109$ voxels.

\[\text{[Insert Table 1 about here]}\]

**Main effects**

The main effect of TASK was tested through the contrast testing for regions with increased neural activity for the UG (compared to the FW) irrespective of the responders’ choice or to the person to which the bargaining was addressed [i.e., $\frac{(URm + UAm + UAt)}{2} – (FWm + FWt)$]. This contrast implicated many regions described playing a critical role in the UG by previous studies, among which bilateral AI, ACC and right DLPFC (Figure 3). Interestingly, the local maxima of these insular, cingulate and prefrontal activations (see Table 1) are always $< 12$ mm distant from the corresponding local maxima reported by Sanfey et al. (2003).

\[\text{[Insert Figure 3 about here]}\]

We further explored effects of RESPONSE within the UG. The analysis of rejections (relative to acceptances), independently of the target of the unfair bargaining, [i.e., $(URm + UAt) – (UAm + UAt)$], implicated the midbrain and the left AI, over and around the left AI cluster isolated through the main effect of TASK (see Figure 4a). Critically, the increased activity for rejection tested in the present contrast was not driven by one target condition only (e.g., Myself), as the insular RESPONSE effect was still significant (albeit at an uncorrected threshold) when considering each target separately ($ts > 2.60$, $ps < 0.05$). No suprathreshold increase of neural activity was associated with acceptances relative to rejections [i.e., $(UAm + UAt) – (URm + UAt)$].
We then tested effects of TARGET and, specifically those increases of neural activity associated with offers addressing oneself (irrespective of whether these were UG or FW) as opposed to offers addressing a third-party [i.e., \((UR_m + UA_m + FW_m) - (UR_t + UA_t + FW_t)\)]. Such increases were found in the ventral part of the medial prefrontal cortex (Figure 4b, violet cluster) and in the right inferior frontal gyrus.

**Interactions**

We first tested for the interaction TASK*TARGET, investigating target-specific increases of neural activity for the UG (relative to FW) task. In particular, we searched for increases of neural activity in the UG which are specific for offers addressing oneself and not the third-party [i.e., \(((UR_m + UA_m)/2 - FW_m) - ((UR_t + UA_t)/2 - FW_t)\)]. The only region surviving correction for multiple comparisons was located in the most anterior portion of the MPFC, around 8 mm above the intercommissural line. Figure 4c (middle graph) displays the parameter estimates extracted from the region’s local maximum, showing an increase of neural activity for Myself (relative to Third-Party) which was limited to UR and UA but not to FW. Critically, this effect was stronger for rejections than acceptances, as revealed by this region exhibiting a significant (albeit only at the uncorrected level) effect also for the contrast \((UR_m - UA_m) - (UR_t - UA_t)\) \(t = 1.68, p < 0.05\). However, the interaction effect isolating MPFC should not be considered a response-bias, as it survived significance also when considering only UG acceptances [i.e., \((UA_m - FW_m) - (UA_t - FW_t)\), \(t = 1.69, p < 0.05\)]. We then tested for regions exhibiting significant BOLD increase for UG (relative to FW), specifically for the third-party condition. We found no suprathreshold effect. Finally, we tested the
RESPONSE*TARGET interaction, thereby assessing target-specific increases of neural activity for specific responses. Also in this case we found no suprathreshold effect.
Discussion

We employed a modified version of the UG (Civai et al., 2010), in which participants played either for themselves (Myself) or on behalf of a third-party. We found the anterior insula involved in dealing with unfair offers affecting both oneself and others. Instead, the middle-anterior portion of the MPFC was recruited exclusively when the unfair offers were related to oneself. Finally, ACC and the right DLPFC were found, at least in our dataset, only broadly involved in the bargaining process (main effect: UG > FW), as their activity was not modulated by the target of the offer or by participants’ choice (but see Supplementary Materials for significant uncorrected difference in right DLPFC activity between one’s and third-party’s rejections). Our data converge with, but also extend, previous studies: we not only mapped the neural mechanisms underlying people’s reaction to unfairness, but we also disentangled those processes reflecting judgments related to unfair behavior per se (fairness), from those related to the emotional consequences of being the victim of the unfair behavior (self effect).

Fairness-Related Neural Networks

Left AI was found active not only when testing effects of UG (as opposed of FW) in both Myself and Third-Party condition, but also when testing rejections (as opposed to acceptances) of UG offers. This result extends what has been found by previous studies (Sanfey et al., 2003; Tabibnia et al., 2008; Güröglu et al., 2010, 2011), by describing left AI activity involved in reacting not only to a self-directed mistreatment, but also to the same mistreatment affecting an unknown other person. Furthermore, our results extend the current understanding about the role played by the insula in UG. Indeed, as previous studies reported this portion of the anterior insula involved in negative
experiences, such disgust, anger, fear, pain or thirst (Damasio et al., 2000; Calder et al., 2001; de Araujo et al., 2003; Wicker et al., 2003; Corradi-Dell’Acqua et al., 2011), it has been argued that this region mediates those negative emotional reactions which, according to the wounded pride/spite model (Pillutla and Murnighan, 1996), favor rejections (Sanfey et al., 2003). Although the wounded pride/spite model has been recently challenged by studies favoring an interpretation of rejections in terms of reinforcement of fairness in the community (Knoch et al., 2006, 2008; Civai et al., 2010; Baumgartner et al., 2011) it still could be argued that being the victim of unfairness triggers a negative emotional reaction, and that the AI involvement in UG rejection is its neural signature. Our data speak against this interpretation and suggest instead that the role played by AI in UG is in reacting to unfairness, irrespective of whether the mistreatment affects participants themselves or an unknown person.

That activity of AI alone cannot be considered evidence of negative emotional arousal was already established by studies associating its activity with positive affect (e.g., Hennenlotter et al., 2005; Jabbi et al., 2007) but also with cognitive processes which are not necessarily emotionally-grounded, such as motor control, memory, attention, etc. (see, Kurth et al., 2010, as meta-analysis). Recent accounts suggest that AI integrates information about modality-specific feelings with cognitive processes, individual preferences and contextual information in order to promote behavioral responses (Singer et al., 2009; Lamm and Singer, 2010). In this perspective this region is an ideal candidate for promoting fairness-related behavior which emerges from the integration of cognitive, emotional and motivational mechanisms (Moll et al., 2008). Indeed, previous studies engaging participants in dyadic social interactions (but not the UG) have suggested that left AI mediates punishments of unfair behavior: for instance, Rilling et al. (2008) implicated coordinates proximal to ours (< 5 mm) in unreciprocated cooperation during the Prisoner’s Dilemma, King-Casas et al. (2008) associated left AI (< 10 mm) with borderlines patients’ inability to maintain
Self versus fairness mechanisms in the ultimatum game

cooperation in a Trust Game, whereas Strobel et al. (2011) reported activations the same region (< 5 mm) when participants sanctioned unfair offers in the Dictator Game. To the best of our knowledge this is the first Ultimatum Game study in which insular activity can be interpreted in terms of sanction of the proposers’ norm violations (but see Gürroğlu et al., 2010 for associating AI to one’s own norm violations). Furthermore, in almost all previous studies using the other economical games, participants’ gain/loss was directly affected by the game’s rules, thus leaving open the possibility that the insular activity they reported was reflective of concerns about one’s welfare. This is not the case of our study in which participants choices in the Third-Party did not affect their own pocket. We therefore believe that our study provides the strongest evidence in favor of AI promoting fairness-related behavior in money bargaining.

Self-Specific Neural Networks

Studies in the field of economics implicated ventral portions of the MPFC in assessing the value of potential outcomes (see, Amodio and Frith, 2006, as review): for instance, the activity of this region was associated by Knutson et al. (2005) with the computation of expected monetary value, and by Coricelli and colleagues (Camille et al., 2004; Coricelli et al., 2005; Larquet et al., 2010) with anticipated regret associated with monetary decision. A similar interpretation of ventral MPFC’s activity is provided by neuropsychological studies using the classical UG: in a first experiment, Koenigs and Tranel (2007) described patients with ventral MPFC damage more prone to reject unfair offers, and interpreted their results as a deficit in emotional control (thus being more exposed to the emotional effects of an unfair treatment); however, in a subsequent experiment, Moretti et al. (2009) replicated Koenigs’ findings, but only when bargaining offers were described as abstract sums to be received later, rather than visible and immediately-available banknotes, thus favoring a deficit in the representation of the offer’s value (the inability to code which, makes
the patients less able to foresee the benefits of accepting). This interpretation of ventral MPFC activity is consistent also with our data which show increased activity in this region whenever participants (but not a third-party) are offered money \((Myself > Third-Party, Figure 4b, violet blobs)\). Furthermore, part of this region exhibited an activity which increased linearly with the amount of money participants gained in the FW task (see Supplementary Materials), thus strengthening the hypothesis of a sensitivity to personal gain rather than in mere self-reflective processing.

A much different interpretation has been offered in literature for the middle-anterior portion of the MPFC (over and above the inter-commissural line) and involves the co-occurrence of cognitive, emotional and social processes (Amodio and Frith, 2006). For instance, the middle-anterior MPFC responds to emotional events (Dolcos et al., 2004; Ochsner et al., 2004; Peelen et al., 2010) and has a signal which correlates with one’s SCR in both gambling tasks and resting state (Patterson et al., 2002). The middle-anterior MPFC has also been implicated also in self-reflection (Zysset et al., 2002; Kelley et al., 2002; Johnson et al., 2002), mentalizing (Goel et al., 1995; Fletcher et al., 1995; Saxe and Powell, 2006) and moral judgments (Greene et al., 2001, 2004; Moll et al., 2002). Amodio et al. (2006) suggested that value-related representations in the ventral MPFC extend the more anterior (and superior), the more complex they become, and that they integrate with socio-affective processes. Our data converge with this distinction: indeed, whereas in our study the ventral MPFC was most likely activated in relation to one’s (but not Third-Party’s) potential gain, the same interpretation cannot be used for the middle-anterior portion of the MPFC associated with the interaction term (Figure 4b, green cluster). Indeed, this latter region showed no differential activation between Myself and Third-Party conditions in the FW, but only during UG acceptances and (more strongly) during rejections. This functional pattern is reminiscent of the one described by Civai et al. (2010) who showed enhanced SCR associated for
Myself (relative to Third-Party) UG especially for rejections. The increase of one’s emotional response in relation to self-directed experimental manipulation converge with recent accounts suggesting that self and affective coding might be instantiated in similar neural networks, as emotional judgments might be considered a self-referential task (Amodio and Frith, 2006) and the Self an emotional entity per se (Modinos et al., 2009). In this perspective the middle-anterior MPFC activity observed in our study might reflect those processes involved in the coding and control of the differential emotional response evoked by being oneself the target of unfairness, especially when this unfairness is subsequently sanctioned at one’s cost. This interpretation of middle-anterior MPFC functioning might also account for results of previous UG studies, such as the modulation of rejection-related activity in this region by the proposer’s intention to be unfair (Güroğlu et al., 2010, although authors offer an interpretation in terms of differential mentalizing). This interpretation is also consistent with recent findings describing the MPFC as part of a network involved in overriding self-interest motives during rejection of unfair UG offers: indeed the activity of this region was found to be affected by transient inactivation of the right DLPFC (Baumgartner et al., 2011) which, in turn, is detrimental for classical UG rejections (van’t Wout et al., 2005; Knoch et al., 2006, 2008; Baumgartner et al., 2011). Interestingly, in our dataset as well the activity of the MPFC and right DLPFC seems coupled (see Supplementary Materials) as both regions show stronger activity when rejecting self-directed (relative to others-directed) offers. Based on previous and present results, it is conceivable that the apparent causal role played by this prefrontal network in promoting rejections would be limited to the case in which the self-interest is relevant, thus not during the third-party UG. Further studies will address this issue.
Conclusion

Rejections in the classical UG can either be interpreted as emotional reactions to a self-directed unfair treatment (“I have been treated unfairly”) or as pure considerations about fairness (“this person has been treated unfairly”) leading to the discouragement of social norm violations. Our data allow this distinction, and show that the anterior insula is specifically involved in fairness-related behavior, whereas the MPFC (and right DLPFC) is involved in monitoring those emotional reactions due to being the direct target of the bargaining.
References


Self versus fairness mechanisms in the ultimatum game


Self versus fairness mechanisms in the ultimatum game


Figure Captions

Figure 1. (A) Four conditions were employed: the first two conditions refer to the UG, whereas the last two conditions refer to the FW. In the first and third conditions, participants’ decisions were related to themselves (Myself trials), whereas in the second and the fourth condition decisions were related to another person (Third-Party trials). (B) Trial Structure.

Figure 2. Behavioral Results. Rejection Rates are plotted as a function of Gain in both UG (black circles) and FW (white triangles) tasks and Myself and Third-Party conditions.

Figure 3. Surface renderings of the functional contrasts testing regions exhibiting a larger neural activity when subjects were engaged in UG, rather than FW.

Figure 4. (A) Sections displaying the functional contrast testing Ultimatum Game > Free Win (blue activation) and the contrast testing Rejections > Acceptances (yellow activation). (B) Sections displaying the functional contrast testing the interaction term (green activations) and Myself > Third-Party (violet activation). (C) The parameter estimates associated with representative voxels of the activated areas are displayed together with 95% confidence intervals (for AI we choose the local maxima obtained when testing Rejection > Acceptances). Red bars refer to offers addressing oneself, whereas cyan bars refer to offers addressing a third-party.
(A) Four conditions were employed: the first two conditions refer to the UG, whereas the last two conditions refer to the FW. In the first and third conditions, participants’ decisions were related to themselves (Myself trials), whereas in the second and the fourth condition decisions were related to another person (Third-Party trials). (B) Trial Structure.

300x374mm (96 x 96 DPI)
Behavioral Results. Rejection Rates are plotted as a function of Gain in both UG (black circles) and FW (white triangles) tasks and Myself and Third-Party conditions.
Surface renderings of the functional contrasts testing regions exhibiting a larger neural activity when subjects were engaged in UG, rather than FW.

46x21mm (300 x 300 DPI)
(A) Sections displaying the functional contrast testing Ultimatum Game > Free Win (blue activation) and the contrast testing Rejections > Acceptances (yellow activation). (B) Sections displaying the functional contrast testing the interaction term (green activations) and Myself > Third-Party (violet activation). (C) The parameter estimates associated with representative voxels of the activated areas are displayed together with 95% confidence intervals (for AI we choose the local maxima obtained when testing Rejection > Acceptances). Red bars refer to offers addressing oneself, whereas cyan bars refer to offers addressing a third-party.

176x212mm (300 x 300 DPI)
Table 1. Voxels showing significant increases of neural activity. All clusters survived correction for multiple comparisons at the cluster level (height threshold $p < 0.001$, uncorrected). Coordinates (in standard MNI space) refer to maximally activated foci. L and R refer to the left hemisphere and right hemisphere, respectively. M refers to medial structures.

<table>
<thead>
<tr>
<th>REGION</th>
<th>SIDE</th>
<th>Coordinates (in standard MNI space)</th>
<th>Cluster Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Main effect of TASK: Ultimatum Game &gt; Free Win</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(URm + UAm + URt + UAt)/2 - (FWm + FWt)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supramarginal Gyrus</td>
<td>R</td>
<td>42</td>
<td>-34</td>
</tr>
<tr>
<td>Supramarginal Gyrus</td>
<td>L</td>
<td>-40</td>
<td>-36</td>
</tr>
<tr>
<td>Calcarine Gyrus</td>
<td>R</td>
<td>10</td>
<td>-62</td>
</tr>
<tr>
<td>Calcarine Gyrus</td>
<td>L</td>
<td>-10</td>
<td>-66</td>
</tr>
<tr>
<td>Precuneus</td>
<td>R</td>
<td>10</td>
<td>-64</td>
</tr>
<tr>
<td>Precuneus</td>
<td>L</td>
<td>-8</td>
<td>-62</td>
</tr>
<tr>
<td>Anterior Insula</td>
<td>R</td>
<td>30</td>
<td>22</td>
</tr>
<tr>
<td>Anterior Insula</td>
<td>L</td>
<td>-34</td>
<td>16</td>
</tr>
<tr>
<td>Midbrain</td>
<td>M</td>
<td>-6</td>
<td>-12</td>
</tr>
<tr>
<td>Anterior Cingulate Cortex</td>
<td>L</td>
<td>-2</td>
<td>18</td>
</tr>
<tr>
<td>Precentral Gyrus</td>
<td>L</td>
<td>-40</td>
<td>-6</td>
</tr>
<tr>
<td>Dorsolateral Prefrontal Cortex</td>
<td>R</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>Middle Temporal Gyrus</td>
<td>L</td>
<td>50</td>
<td>-62</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>-30</td>
<td>-78</td>
</tr>
<tr>
<td>Inferior Occipital Gyrus</td>
<td>L</td>
<td>36</td>
<td>-84</td>
</tr>
<tr>
<td>Inferior Frontal Gyrus</td>
<td>R</td>
<td>48</td>
<td>4</td>
</tr>
</tbody>
</table>
Main effect of RESPONSE: Rejected > Accepted Ultimatum Game offers
\[(UR_m + UR_t) - (UA_m + UA_t)\]

<table>
<thead>
<tr>
<th>Structure</th>
<th>M</th>
<th>L</th>
<th>Value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midbrain/PAG</td>
<td>-6</td>
<td>-36</td>
<td>-26</td>
<td>-4</td>
</tr>
<tr>
<td>Anterior Insula</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Main effect of TARGET: Myself > Third-Party
\[(UR_m + UA_m + FW_m) - (UR_t + UA_t + FW_t)\]

<table>
<thead>
<tr>
<th>Structure</th>
<th>M</th>
<th>L</th>
<th>Value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medial prefrontal cortex (ventral)</td>
<td>-2</td>
<td>48</td>
<td>38</td>
<td>-6</td>
</tr>
<tr>
<td>Inferior Frontal Gyrus</td>
<td></td>
<td>26</td>
<td>46</td>
<td>-6</td>
</tr>
</tbody>
</table>

**TASK*TARGET interaction: Ultimatum Game > Free Win, specifically for Myself**
\[((UR_m + UA_m)/2 - FW_m) - ((UR_t + UA_t)/2 - FW_t)\]

<table>
<thead>
<tr>
<th>Structure</th>
<th>M</th>
<th>L</th>
<th>Value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medial prefrontal cortex (middle-anterior)</td>
<td>0</td>
<td>26</td>
<td>58</td>
<td>8</td>
</tr>
</tbody>
</table>

* p < 0.001; \(^\dagger\) p < 0.01; \(^\ddagger\) p < 0.05, corrected for multiple comparisons for the whole brain.
\(^\$\) p < 0.05, corrected for the small volume