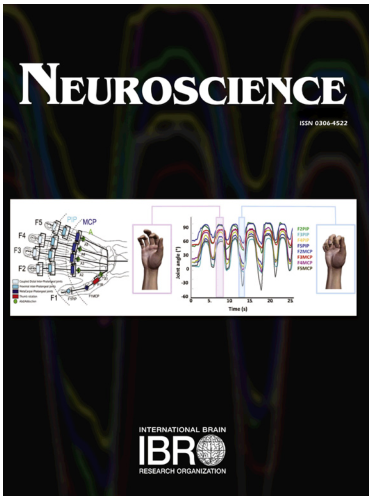
Accepted Manuscript



Anterior insula signals inequalities in a modified Ultimatum Game

Xuemei Cheng, Li Zheng, Lin Li, Yijie Zheng, Xiuyan Guo, Guang Yang

PII: S0306-4522(17)30106-9
DOI: http://dx.doi.org/10.1016/j.neuroscience.2017.02.023
Reference: NSC 17612

To appear in: *Neuroscience*

Received Date: 8 November 2016
Revised Date: 9 February 2017
Accepted Date: 13 February 2017

Please cite this article as: X. Cheng, L. Zheng, L. Li, Y. Zheng, X. Guo, G. Yang, Anterior insula signals inequalities in a modified Ultimatum Game, *Neuroscience* (2017), doi: http://dx.doi.org/10.1016/j.neuroscience.2017.02.023

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Anterior insula signals inequalities in a modified Ultimatum Game

Xuemei Cheng[a], Li Zheng[b], Lin Li[c], Yijie Zheng[a], Xiuyan Guo[b,CA], Guang Yang[a]

[a]Shanghai Key Laboratory of Magnetic Resonance, Department of Physics, East China Normal University, Shanghai, China

[b] Shanghai Key Laboratory of Magnetic Resonance and Key Laboratory of Brain Functional Genomics, Ministry of Education, Shanghai Key Laboratory of Brain Functional Genomics, East China Normal University, Shanghai, China

[c]School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

[CA]Corresponding Author and Address:
Dr. Xiuyan Guo
Address: School of Psychology and Cognitive Science, East China Normal University, North Zhongshan Road 3663, Shanghai, SH, 200062, China
E-mail: xyguo@psy.ecnu.edu.cn
Phone: 086-021-62232908

Abstract

Studies employing the Ultimatum Game (UG) which involves two parties (i.e., proposers and responders) splitting some money have suggested the role that anterior insula (AI) plays in detecting fairness norm violation, i.e., violation of the responder’s expectation of receiving equal splits from the proposer. In this study, we explored how AI would respond when there existed simultaneously another expectation of being treated equivalently as others. Participants acted as responders and would be informed about both the offers they received and the average amount of money the same proposer offered to others. Hence we introduced different conditions where participants were treated equivalently or not equivalently as other responders in UG. Participants could decide to accept or reject the offer with acceptance leading to the suggested split and rejection leaving both parties nothing. Behavioral results showed that participants rejected more unfair offers and reacted more slowly during acceptance (vs. rejection) of offers when they were offered less than others. At the neural level, stronger AI activation was observed when participants received unfair relative to fair offers, as well as when they received unequal relative to equal offers. Moreover, dorsomedial prefrontal cortex/dorsal anterior cingulate cortex (dmPFC/dACC) exhibited greater activity during receiving unequal (vs. equal) offers and during acceptance (vs. rejection) of offers which were less than others’. Taken together, the present study demonstrated that the treatment of others modulated both behavioral responses to unfairness and neural correlates of the fairness-related decision-making process, and that AI played a general role in detecting norm violations.

Key words
anterior insula; unfairness; norm violation; expectation; fMRI.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Introduction

Shared expectations on how people should behave in a given circumstance constitute social norms in human society (Hechter and Opp, 2001;Xiang et al., 2013). The fairness norm is based on the expectation that everyone should be treated equally when everything else is the same (Bohnet and Zeckhauser, 2004;Civai et al., 2012). When such norm is violated, individuals would rather punish norm violations even at their own cost. For instance，in the widely studied Ultimatum Game (UG), a responder will reject an unequal division of an amount of money which is made by a proposer, even though rejection means that both the responder and the proposer would leave with nothing and that acceptance makes them get the money as suggested (Güth et al., 1982;Camerer and Thaler, 1995). The rejection behavior which contradicts the standard notions of money-maximizing preferences reflects that people care about fairness and are sensitive to norm violations.

Anterior insula (AI) has been suggested as a key brain region associated with detecting norm violations by evidence of its involvement in signaling deviations from people’s expectations (Montague and Lohrenz, 2007;Spitzer et al., 2007;King-Casas et al., 2008;Chang and Sanfey, 2013;Xiang et al., 2013). Abundant neuroimaging studies employing UG have demonstrated that AI was strongly activated when people received unfair relative to fair offers (Sanfey et al., 2003;Guo et al., 2013;Guo et al., 2014;Cheng et al., 2015). The traditional UG paradigm typically involves two interacted parties and people who act as the responder would automatically take the proposer’s income as reference and expect equal splits (Bohnet and Zeckhauser, 2004;Wu et al., 2011;Civai et al., 2012). Thus, the heightened level of AI activation during receiving unfair offers was associated with violation of the expectation of receiving equal splits. Note that AI was not only sensitive to disadvantageous inequalities caused by offers less than 50% of the total amount, even advantageous inequalities (i.e., offers more than 50% of the total amount) activated AI (Civai et al., 2012). To sum up, unequal splits between the proposer and the responder violate people’s expectation and AI plays an important role in detecting the deviation from people’s expectations.

Apart from expecting equal splits between proposers and themselves, numerous studies have demonstrated that people also care about others in similar situations and expect to be treated equivalently as them (Babcock et al., 1996;Bohnet and Zeckhauser, 2004;Wu et al., 2011). It is of great interest to understand how people would deal with both kinds of expectation violations when they simultaneously exist in a single event, and how AI would respond in such a case. Wu et al. (2011) and Zheng et al. (2015) introduced the treatment of others in UG by informing participants about both the proposals they received and the proposals between other proposers and responders, to investigate how these two expectations affected people’s behavioral and neural responses to unfairness. However, these studies were not able to provide complete answers to the questions mentioned above. The ERP-study of Wu et al. (2011) did not mention anything about AI, which might be due to the difficulties in measuring signals in AI. The fMRI study of Zheng et al. (2015) focused on the situation where people were offered less than others and found significant increase in AI activity compared with the situation where people were offered equally to others. But the study did not investigate another case of violating the expectation of being offered equally to others, that is to say, being offered more instead of less than others. Thus, further research is still in need.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

In the present study, we modified the traditional UG paradigm to investigate how AI would respond when two expectations simultaneously existed, i.e., the expectation of receiving equal splits from the proposer and the expectation of being treated equivalently as others. Participants acted as the responder and received fair or unfair proposals about how to split ￥50 from unknown proposers. On this basis, in order to introduce the treatment of other responders, the average amount of money the same proposer offered to other responders was also presented. Participants were offered￥5 more than the average offer in the *MoreAve* condition and were offered￥5 less than the average offer in the *LessAve* condition. In the *EqualAve* condition, participants received offers equal to the average offer. Based on previous findings which revealed increased rejection rates when people were offered less than others comparing with being offered either more than or equal to others (Wu et al., 2011;Zheng et al., 2015), we predict that higher rejection rates would be observed in the *LessAve* condition compared with the *MoreAve* and the *EqualAve* conditions. At the neural level, increased AI activity was expected when participants received unfair offers which violated the fairness norm. Moreover, considering that offers unequal to what others received would also violate participants' expectation, heightened level of AI activation was expected in both *MoreAve* and *LessAve* conditions compared with the *EqualAve* condition.

Apart from AI, there is also empirical evidence demonstrating that dorsal anterior cingulate cortex (dACC) is associated with detecting expectation violation in UG (Chang and Sanfey, 2013). Thus, the treatment of others might also modulate the involvement of dACC.

Experimental Procedures

Participants

Twenty right-handed volunteers [10 females, mean age = 23.8 ± 1.9 (s.d.) years] took part in this study. All the participants had normal or corrected-to-normal vision and reported no abnormal neurological history. One participant was excluded from further statistical analyses due to severe head motion (> 3°) during scanning. Written informed consent was acquired from all the participants. The study was approved by the Ethics Committee of East China Normal University.

Materials

104 face pictures (52 depicting females) were selected from Chinese Facial Affective Picture System (Gong et al., 2011) and were displayed as proposers. These pictures were randomly allocated to 3 conditions (*Treatment*: *MoreAve*, *EqualAve* and *LessAve*). There were 40 pictures in the *EqualAve* condition [8 pictures for each of five different proposals (￥5:￥45,￥10:￥40, ￥15:￥35, ￥20:￥30 and ￥25:￥25)], 32 pictures in the *MoreAve* condition [(8 pictures for each of four different proposals (￥10:￥40, ￥15:￥35, ￥20:￥30 and ￥25:￥25)] and 32 pictures in the *LessAve* condition [(8 pictures for each of four different proposals (￥5:￥45, ￥10:￥40, ￥15:￥35 and ￥20:￥30)]. Each condition had an equal number of male and female face pictures. The emotion valence, arousal and attractiveness of pictures were counterbalanced across different conditions.

Procedure

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Participants were told that they would participate in an economic game with different partners, along with an instruction introducing the rule of the game. They were told that they would be presented with a proposer's proposal about how to split ￥50 between them. In addition, the average amount of money the same proposer offered to other responders would also be presented. They were told that their offers could be equal to the average offer, or lower or higher than that. Then they were told that they could decide to either accept or reject the proposal with acceptance leading to the suggested split and rejection leaving both of them nothing. Participants were also informed that the proposals were obtained from different proposers before the experiment. As for the payment, participants were told that several trials would be randomly selected and that both themselves and the proposers would be paid according to their decisions. Finally, 5% of the total trials (5 trials) were selected and participants were paid accordingly. Additionally, each participant was also paid ￥50 bonus for taking part in this experiment.

Before scanning, participants practiced 26 trials on a laptop. After the practice, participants completed 104 trials in the scanner. There were 40 trials and 5 kinds of proposals (￥5:￥45,￥10:￥40, ￥15:￥35, ￥20:￥30 and ￥25:￥25) in the *EqualAve* condition, 32 trials and 4 kinds of proposals (￥10:￥40, ￥15:￥35, ￥20:￥30 and ￥25:￥25) in the *MoreAve* condition, and 32 trials and 4 kinds of proposals (￥5:￥45, ￥10:￥40, ￥15:￥35 and ￥20:￥30) in the *LessAve* condition, with each kind of proposals having 8 trials. The amount of money participants received was equal to the average offer in the *EqualAve* condition, ￥5 more than the average offer in the *MoreAve* condition and ￥5 less than the average offer in the *LessAve* condition. Considering that it would be unaccountable for participants if they knew the average offer was ￥0 or larger than half-split, there were no proposals of ￥5:￥45 in the *MoreAve* condition (i.e., the average offer would be ￥0) and no proposals of ￥25:￥25 in the *LessAve* condition (i.e., the average offer would be more than￥25). All of the trials were presented in a random order. For each trial, the proposal screen was presented for 6s indicating both the split between the proposer and the participant and the average amount of money the same proposer offered to others. Then a decision cue appeared and participants were required to decide whether to accept or reject the offer within 3s by pressing corresponding buttons of the magnet-compatible button Box (i.e., right index finger for acceptance and right middle finger for rejection). Once they responded, a blue frame outside the selected choice would be presented for 1s to provide participants with the feedback of their decision. The intervals between trials were jittered from 1 to 4 s. There was also one jittered blank (500~1100 ms) between the proposal screen and the decision cue (**Figure 1**).

After scanning, the same stimuli including proposers' proposals and their corresponding average offers were presented again. Participants were asked to rate the fairness of each offer on a 9-point Likert-type scale with 1 indicating extremely unfair and 9 indicating extremely fair.

Behavioral data Analysis

Statistical analyses on rejection rates, fairness ratings and reaction times (RTs) were performed for each kind of proposals separately with responded trials. Trials which participants failed to respond to were excluded from any further analyses [mean = 1.00 ± 1.37 (s.d.) trials]. For proposals like ￥5:￥45 and ￥25:￥25, which only appeared on two *Treatment* levels, paired *t*-tests were conducted. For proposals like ￥10:￥40, ￥15:￥35 and ￥20:￥30, which appeared on all

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

three *Treatment* levels, one-way repeated measures ANOVAs were performed. Holm’s sequential Bonferroni correction was used to correct for multiple comparisons (Holm, 1979). If any significant result was observed in ANOVAs, *post hoc* pairwise comparisons were conducted and multiple comparisons were corrected by using Holm’s sequential Bonferroni correction again.

For rejection rates, statistical analyses were performed as above except for proposals of ￥25:￥25 and proposals of ￥20:￥30 since no rejection was observed for proposals of ￥25:￥25 and for proposals of ￥20:￥30 in the *MoreAve* and the *EqualAve* conditions. Instead, a one-sample *t*-test was conducted on rejection rates of proposals of ￥20:￥30 in the *LessAve* condition. It is worth noting that logistic regression should be a better way to analyze the binary rejection rate data. However, the unitary response observed in some conditions (e.g. no rejection for proposals of ￥20:￥30 in the *MoreAve* or the *EqualAve* conditions) made it impossible to estimate parameters in the logistic regression model. Thus, ANOVAs were still used in the present study.

fMRI Image Acquisition and Analysis

Participants were scanned using a 3T Siemens scanner at the Shanghai Key Laboratory of Magnetic Resonance of East China Normal University. We first acquired each participant’s anatomical images using a T1-weighted, multiplanar reconstruction (MPR) sequence (TR = 2530 ms, TE = 3.42 ms, 192 slices, slice thickness = 1 mm, FOV = 256 mm, matrix size = 256 ∗ 256) (Cheng et al., 2015;Wang et al., 2015). After that, functional images were acquired using a gradient-echo echo-planar imaging (EPI) sequence (TR = 2200 ms, TE = 30 ms, FOV = 220 mm, matrix size = 64 ∗ 64, 35 slices, slice thickness = 3 mm, gap = 0.3 mm) (Cheng et al., 2015;Wang et al., 2015).

The preprocessing and statistical analyses of brain imaging data were performed with the SPM8 software package (Wellcome Department of Cognitive Neurology, London). The first five functional images were discarded from each subject to allow scanner equilibrium effects. Then, all functional images were slice timing corrected, realigned, normalized into the MNI space (resampled at 2 mm ∗ 2 mm ∗ 2 mm voxels), and smoothed with an 8-mm full-width half maximum isotropic Gaussian kernel (Cheng et al., 2015).

First-level analyses were then performed across the whole brain for each subject using two general linear models (GLM). The norm-related model was used to explore neural responses to violations of two kinds of expectations, i.e., the expectation of receiving equal splits from the proposer and the expectation of being treated equivalently as others. The response-related model accounted for the modulation of others’ treatment on participants’ neural correlates of the fairness-related decision-making process. In the norm-related model, onsets of the proposal screen and onsets of the decision cue were modeled for five types of events, including $Fair_{MoreAve}$ (fair offers in the *MoreAve* condition), $Unfair_{MoreAve}$ (unfair offers in the *MoreAve* condition), $Fair_{EqualAve}$ (fair offers in the *EqualAve* condition), $Unfair_{EqualAve}$ (unfair offers in the *EqualAve* condition) and $Unfair_{LessAve}$ (unfair offers in the *LessAve* condition). In the response-related model, we modeled onsets of the proposal screen and onsets of the decision cue for six types of events, including $Acc_{MoreAve}$ (accepted offers in the *MoreAve* condition), $Rej_{MoreAve}$ (rejected offers in the *MoreAve*

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

condition), *Acc*$_{EqualAve}$ (accepted offers in the *EqualAve* condition), *Rej*$_{EqualAve}$ (rejected offers in the *EqualAve* condition), *Acc*$_{LessAve}$ (accepted offers in the *LessAve* condition) and *Rej*$_{LessAve}$ (rejected offers in the *LessAve* condition). One participant had to be excluded from the response-related model due to lack of efficient number of trials (less than 5 trials) for acceptance in the *LessAve* condition. Additionally regressors of no interest in both two models were the feedback for acceptance, the feedback for rejection, and proposal screen and decision cue for trials which participants failed to respond to. All these regressors were modeled with zero duration and convolved with a canonical hemodynamic response function (HRF). Moreover, six realignment parameters and one overall mean during the whole phase were included in the design matrix as well. To filter the low-frequency noise, a cutoff of 128s was applied. During first-level analyses, five contrast images (*Fair*$_{MoreAve}$, *Unfair*$_{MoreAve}$, *Fair*$_{EqualAve}$, *Unfair*$_{EqualAve}$ and *Unfair*$_{LessAve}$) for proposal presentation were acquired from each participant in the norm-related model. In the second-level analyses, these images were fed into a flexible design which employed a random effects model. For the response-related model, six contrast images (*Acc*$_{MoreAve}$, *Rej*$_{MoreAve}$, *Acc*$_{EqualAve}$, *Rej*$_{EqualAve}$, *Acc*$_{LessAve}$ and *Rej*$_{LessAve}$) for proposal presentation were acquired from each participant and were fed into another flexible design in the second-level analyses.

For the norm-related model, we first explored neural responses to one expectation violation by fixing the status of the other expectation. Specifically, the *(Unfair - Fair)*$_{EqualAve}$ and the reverse contrasts were conducted to investigate participants' neural responses to offers which were unfair but equal to what others received. In order to investigate brain activities associated with receiving unequal treatment compared with others, an *F*-contrast (*Unfair*$_{MoreAve}$, *Unfair*$_{EqualAve}$ and *Unfair*$_{LessAve}$) which only used unfair trials was conducted, since all the three levels of *Treatment* only simultaneously existed when participants received unfair offers. Then, we explored brain activities related to both expectation violations. The five types of events were sorted into three conditions. The *Both*$_{Violated}$ condition consisted of *Unfair*$_{MoreAve}$ and *Unfair*$_{LessAve}$ in which both two expectations were violated. The *Single*$_{Violated}$ condition included *Fair*$_{MoreAve}$ and *Unfair*$_{EqualAve}$ in which only one expectation was violated. The *Non*$_{Violated}$ condition involved only *Fair*$_{EqualAve}$ in which neither expectation was violated. The *(Both - Single)*$_{Violated}$ contrast, the *(Single - Non)*$_{Violated}$ contrast and their corresponding reverse contrasts were computed. As for the response-related model, an *F*-contrast was computed first to examine the main effect of *Treatment* (*MoreAve*, *EqualAve* and *LessAve*). Then, the (*Reject - Accept*) contrast and the reverse contrast were performed to examine the main effect of *Response*. The interaction between *Treatment* (*MoreAve*, *EqualAve* and *LessAve*) and *Response* (*Accept* vs. *Reject*) was examined by another *F*-contrast. For all the analyses, we only reported activations which surviving the voxel-level threshold of Family-wise error (FWE) corrected $p < 0.05$ with an extent threshold of 10 voxels, unless otherwise stated. The MarsBaR toolbox was used to extract percentage signal change when significant activations were observed.

Moreover, in order to see whether the activation patterns observed in the above analyses still remained after controlling for the effect of RT, additional analyses including trial-by-trial RT across all conditions as a parametric modulator were conducted. Specifically, a regressor consisting of all onsets of the proposal screen regardless of event types and multiple modulators with one for the RT data and others coding different event types (i.e., five event types in the

ACCEPTED MANUSCRIPT

norm-related model and six event-types in the response-related model) were specified in the new GLMs. Additional regressors and the high-pass filter cutoff were the same as those in the initial two first-level GLMs in which RT was not controlled for. Then, five contrast images ($Fair_{MoreAve}$, $Unfair_{MoreAve}$, $Fair_{EqualAve}$, $Unfair_{EqualAve}$ and $Unfair_{LessAve}$) in the norm-related model and six contrast images ($Acc_{MoreAve}$, $Rej_{MoreAve}$, $Acc_{EqualAve}$, $Rej_{EqualAve}$, $Acc_{LessAve}$ and $Rej_{LessAve}$) in the response-related model for proposal presentation were acquired from each participant. These images were fed into the same flexible designs as those when RT was not controlled for, with the same contrasts being conducted as well.

Results

Behavioral results

Statistical analyses on rejection rates (**Figure 2A**) only revealed a significant main effect of *Treatment* on proposals of ￥15:￥35 [$F(2, 36) = 29.72$, $p < 0.01$]. *Post hoc* pairwise comparisons showed that proposals of ￥15:￥35 were more often rejected in the *LessAve* condition than either the *MoreAve* or the *EqualAve* condition (*ps* < 0.01).

As for fairness ratings (**Figure 2B**), significant main effects of *Treatment* were observed on proposals of ￥10:￥40, ￥15:￥35 and ￥20:￥30 (*Fs* > 10.01, *ps* < 0.01). *Post hoc* pairwise comparisons showed that fairness ratings in the *MoreAve* and the *EqualAve* conditions were higher than those in the *LessAve* condition (*ps* < 0.01) for all these three kinds of proposals. For proposals of ￥25:￥25, fairness ratings were found to be higher in the *EqualAve* condition compared with the *MoreAve* condition ($t = 2.58$, $p < 0.05$).

Statistical analyses on RTs (**Figure 2C**) only revealed a significant main effect of *Treatment* on proposals of ￥20:￥30 [$F(2, 36) = 6.61$, $p < 0.05$]. *Post hoc* pairwise comparisons showed that participants reacted more slowly in the *LessAve* condition compared with either the *MoreAve* or the *EqualAve* condition (*ps* < 0.05). Additionally, we conducted a 3 (*Treatment*: *MoreAve*, *EqualAve* and *LessAve*) ∗ 2 (*Response*: *Accept* vs. *Reject* repeated measures ANOVA on RTs (**Figure 2D**) according to the response-related model in the fMRI analyses. One participant was excluded from the response-related model due to lack of inefficient number of trials during acceptance in the *LessAve* condition. Thus, RTs from 18 participants were analyzed in the repeated measures ANOVA. The results revealed a significant main effect of *Treatment* [$F(2, 34) = 6.52$, $p < 0.01$]. *Post hoc* pairwise comparisons showed that RTs in the *MoreAve* and the *LessAve* conditions were longer than those in the *EqualAve* condition (*ps* < 0.05). Moreover, the interaction between *Treatment* and *Response* was also significant [($F(2,34) = 4.33$, $p < 0.05$). Longer RTs were observed during acceptance relative to rejection of proposals in the *LessAve* condition ($p < 0.05$) but not in the *MoreAve* or the *EqualAve* condition (*ps* > 0.1).

fMRI results

Norm-related model

Increased left AI (MNI -24 24 -2) activity was observed in the $(Unfair - Fair)_{EqualAve}$ contrast (**Figure 3A**), indicating that AI was more strongly activated when participants received unfair offers which were equal to what others received. Meanwhile, the *F*-contrast ($Unfair_{MoreAve}$, $Unfair_{EqualAve}$ and $Unfair_{LessAve}$) also revealed significant activation in left AI (MNI -32 22 -4,

ACCEPTED MANUSCRIPT

Figure 3B). Further analyses on percentage signal change showed that left AI activity was stronger in the *MoreAve* and the *LessAve* conditions compared with the *EqualAve* condition ($ps < 0.05$), while no significant activity difference was found between the *MoreAve* and the *LessAve* conditions ($p > 0.1$). No other suprathreshold activations were detected in these two contrasts or in the *(Fair - Unfair)*$_{EqualAve}$ contrast.

Left AI (MNI -30 20 -2) was significantly activated in the *(Both - Single)*$_{Violated}$ contrast (**Figure 3C**). Similar left AI (MNI -26 24 -2) and right AI (MNI 30 28 -4) activations were revealed in the *(Single - Non)*$_{Violated}$ contrast (**Figure 3D**). Significant right lingual gyrus (MNI 8 -70 4) activation was also detected in the *(Single - Non)*$_{Violated}$ contrast. No suprathreshold activations were detected in the reverse contrasts.

After controlling for RT, similar significant AI activations were still observed in the *(Unfair - Fair)*$_{EqualAve}$ contrast (MNI -24 24 -2 and MNI 32 30 -4), in the *F*-contrast (*Unfair*$_{MoreAve}$, *Unfair*$_{EqualAve}$ and *Unfair*$_{LessAve}$) (MNI -32 22 -4, voxel-level FWE corrected $p < 0.05$), in the *(Both - Single)*$_{Violated}$ contrast (MNI -30 20 -2) and in the *(Single - Non)*$_{Violated}$ contrast (MNI -26 24 -2 and MNI 30 28 -4). Additional significantly activated brain areas included right calcarine sulcus (MNI 10 -72 8) detected in the *(Unfair - Fair)*$_{EqualAve}$ contrast and right lingual gyrus (MNI 8 -70 6) and left superior parietal lobule (MNI -30 -60 48) detected in the *(Single - Non)*$_{Violated}$ contrast.

Response-related model

The *F*-contrast which examined the main effect of *Treatment* revealed significant activations in bilateral AI (MNI -32 22 -4 and MNI 34 22 -8) and dorsomedial prefrontal cortex (dmPFC)/dACC (MNI 4 28 40) (**Figure 4A**). Additional activated brain areas were listed in **Table 1**. As can be seen in **Figure 4A**, activations in left AI and dmPFC/dACC were stronger in the *MoreAve* and the *LessAve* conditions compared with the *EqualAve* condition ($ps < 0.05$), whereas no significant activation difference was found between the *MoreAve* and the *LessAve* conditions ($ps > 0.1$). Right AI was also found to be more active in the *LessAve* condition compared with the *EqualAve* condition ($p < 0.05$) but only showed a trend of activity difference between the *MoreAve* and the *EqualAve* conditions ($p > 0.05$). The *(Reject - Accept)* contrast revealed no significant activations of interest and the reverse contrast revealed no suprathreshold activations (**Table 1**).

The *F*-contrast testing the interaction effect between *Treatment* and *Response* revealed significant activation in dmPFC/dACC (MNI 0 30 40, **Figure 4B**). Additional activated brain regions were listed in **Table 1**. As shown in **Figure 4B**, analyses on percentage signal change showed that dmPFC/dACC was more active during acceptance relative to rejection of offers in the *LessAve* condition ($p < 0.05$) but not in the *MoreAve* or the *EqualAve* condition ($ps > 0.1$).

Table 1 Regions showing Main Effects and Interaction in the Response-related Model

	Peak Activation				
Region	X	Y	Z	*F*/*t*-Value	Voxels
***F-contrast: Main Effect of Treatment* (MoreAve, EqualAve and LessAve)**					

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

L	AI	-32	22	-4	40.39	308
R		34	22	-8	26.94	121
	dmPFC/dACC	4	28	40	35.63	786
L	Rolandic Operculum	-42	-32	24	26.09	44
R	Thalamus	6	-16	0	24.58	32
***t-contrast: Main Effect of Response** (Reject - Accept)*						
R	Medial Superior Frontal Gyrus	6	62	26	6.67	157
L	Posterior Cingulate Cortex	-12	-42	10	5.99	14
***t-contrast: Main Effect of Response** (Accept - Reject)*						
	No Regions					
F-contrast: Treatment** (MoreAve, EqualAve and LessAve)* ∗ ***Response** (Accept vs. Reject)* ***Interaction						
	dmPFC/dACC	0	30	40	21.25	96
R	Superior Frontal Gyrus	24	64	18	18.71	11
R	Supplementary Motor Area	12	20	58	17.99	12

Note. Coordinates (mm) are in MNI space. L = left hemisphere; R = right hemisphere.

All clusters reported survived the voxel-level threshold of Family-wise error (FWE) corrected p < 0.05 with an extent threshold of 10 voxels.

After controlling for RT, similar significant activations of AI (MNI -32 22 -4 and MNI 36 24 -6), dmPFC/dACC (MNI 6 30 42), left rolandic operculum (MNI -42 -32 24), right thalamus (MNI 8 -18 0) and left inferior frontal gyrus (MNI -46 20 4) were observed in the *F*-contrast testing the main effect of *Treatment*. The *(Reject - Accept)* contrast also revealed similar significant activations in right medial superior frontal gyrus (MNI 6 64 26) and left posterior cingulate cortex (MNI -12 -42 10). As for the *F*-contrast testing the interaction effect between *Treatment* and *Response*, similar right superior frontal gyrus (MNI 22 66 18) and dmPFC/dACC (MNI -2 34 38 and MNI 12 30 42, voxel-level FWE corrected p < 0.05) activities were observed.

Discussion

In the present study, we used a modified version of UG, in which participants were informed about not only the offer he/she received but also the average offer the same proposer offered to others, to further explore how AI would respond to norm violations when the expectation of receiving equal splits and the expectation of being treated equivalently as others simultaneously existed. Previous studies have investigated the influence of expectation on fairness-related decision-making. Chang and Sanfey (2013) elicited participants' initial expectation by asking them the offers they believed to receive. In the study of Xiang et al. (2013), researchers manipulated participants' expectations by training them to be preadapted to high/low offers and then giving them medium offers, or vice versa. Expanding on these studies, our study explored the situation where there

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

(Hechter and Opp, 2001)existed simultaneously two expectations by introducing the treatment of others in UG. At the behavioral level, treatment of others exhibited influence on rejection rates of unfair proposals (e.g., ￥15:￥35), indicated by higher rejection rates in the *LessAve* condition compared with either the *MoreAve* or the *EqualAve* condition. Furthermore, longer RTs were observed during acceptance relative to rejection of offers in the *LessAve* condition, indicating that acceptance was a more time consuming choice when participants were offered less than others.

At the neural level, stronger AI activation was revealed when participants received offers which were unfair but equal to the average offer others received, confirming the sensitivity of AI to fairness norm violation (Sanfey et al., 2003;Guo et al., 2013;Guo et al., 2014;Cheng et al., 2015). Furthermore, the treatment of others showed a V-shaped modulation on AI activity. Specifically, increased AI activity was observed during both receiving more and receiving less amount of money than others when compared with receiving the same amount of money as others, whereas no significant activity difference was detected between receiving more and receiving less money than others. These results suggested that AI was sensitive to deviation from either the expectation of receiving equal splits or the expectation of being treated equivalently as others. In addition, we also observed highest level of AI activity when both expectations were violated, followed by situations in which only one of the two expectations was violated, and least AI activity when no expectation was violated. Taken together, these results might provide further evidence that AI plays a general role in detecting multiple norm violations (Montague and Lohrenz, 2007;Civai et al., 2012;Chang and Sanfey, 2013;Corradi-Dell'Acqua et al., 2013;Xiang et al., 2013). Another thing has to be mentioned was that, though some previous studies revealed the association between rejection rates and AI activation in UG (Sanfey et al., 2003;Tabibnia et al., 2008), some other studies observed increased AI activity without increased rejection rates (Baumgartner et al., 2011;Civai et al., 2012;Xiang et al., 2013). Also, in one of our previous studies employing UG (Cheng et al., 2015), we observed a significant decrease in rejection rates but not in AI activation when participants did not (vs. did) have the power to punish proposers. In the present study, stronger AI activity was observed in both *MoreAve* and *LessAve* conditions, whereas only in the *LessAve* condition were higher rejection rates observed. These results suggested that although AI played an important role in detecting norm violations, it might not necessarily predict rejection rates in UG. The rejection behavior in UG might be a consequence after integrating the perception of norm violation and other contextual information.

Moreover, dmPFC/dACC was found to be more active in the *MoreAve* and the *LessAve* conditions relative to the *EqualAve* condition, showing a similar activity pattern to AI, which provided support that dmPFC/dACC might be associated with expectation violation (Chang and Sanfey, 2013). In addition, stronger activation of dmPFC/dACC was observed during acceptance relative to rejection of offers in the *LessAve* condition but not in the *MoreAve* or the *EqualAve* condition. Previous studies employing UG have also suggested that dmPFC/dACC might be associated with conflicts between acceptance and rejection decisions (Sanfey et al., 2003;Gabay et al., 2014;Feng et al., 2015). Also, there is abundant evidence demonstrating that dmPFC/dACC plays an important role in detecting response conflicts (Botvinick et al., 1999;Botvinick et al., 2004;Kerns et al., 2004;Ridderinkhof et al., 2004;Amodio and Frith, 2006). Based on the findings that participants were more tended to reject the proposal when they were offered less than others

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

(Bohnet and Zeckhauser, 2004;Wu et al., 2011;Zheng et al., 2015), to accept offers less than what others received might be a tough decision and involved more conflicts. Thus, the increased dmPFC/dACC activity observed during acceptance relative to rejection of offers which were less than others' might be associated with increased conflicts during making the tough decision of acceptance. The RT data also showed a similar morphology in comparison to the activity pattern of dmPFC/dACC. However, when we included the trial-by-trial RT as a parametric regressor (across conditions) for both of the two first-level GLMs to control for the effect of RT, dmPFC/dACC activity patterns were still similar to those when RT was not controlled for. This might be due to relatively low reliability of RT data in the present study since they were recorded at the decision phase, which was more than 6 seconds after the onset of the proposal. Our further research would consider recording RTs at the phase of proposal presentation to make the RT data more effective.

In conclusion, in the present study, we further explored how AI would respond to two kinds of expectation violations by adopting a modified UG paradigm in which responders were informed about not only the offers they received but also the average amount of money the same proposer offered to others. Behaviorally, we found that people rejected unfair offers more often when they were offered less than others. Meanwhile, reaction times were longer for accepting than rejecting such offers. At the neural level, AI was found to be more active both when people received unfair (vs. fair) offers and when they were offered more or less (vs. equal) than others, indicating that AI was sensitive to deviations from either the expectation of receiving equal splits from the proposer or the expectation of being treated equivalently as others. Furthermore, AI exhibited highest level of activation when both two expectations were violated, followed by the situations in which only one expectation was violated, and the least when no expectation was violated. These results suggested that AI played a general role in detecting norm violations. Moreover, increased dmPFC/dACC activity was observed when people were offered unequally (vs. equally) to others and during acceptance relative to rejection of offers which were less than others'. Taken together, our study further demonstrated that the treatment of others affected people's both behavioral responses to unfairness and neural correlates of the fairness-related decision-making process, as well as the important role of AI in detecting norm violations.

Acknowledgements
This research was supported by the National Natural Science Foundation of China (31271090; 71371180), the Key Program of the National Social Science Foundation of China (14AZD106) and Shanghai Leading talent Plan.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

References

Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7:268-277.

Babcock L, Wang X, Loewenstein G (1996) Choosing the Wrong Pond: Social Comparisons That Reflect a Self-Serving Bias. Quart J Econ 111:1-19.

Baumgartner T, Knoch D, Hotz P, Eisenegger C, Fehr E (2011) Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. Nat Neurosci 14:1468-1474.

Bohnet I, Zeckhauser R (2004) Social Comparisons in Ultimatum Bargaining. Scand J Econ 106:495–510.

Botvinick M, Nystrom LE, Fissell K, Carter CS, Cohen JD (1999) Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature 402:179-181.

Botvinick MM, Cohen JD, Carter CS (2004) Conflict monitoring and anterior cingulate cortex: an update. Trends in cognitive sciences 8:539-546.

Camerer C, Thaler RH (1995) Anomalies: Ultimatums, dictators and manners. J Econ Perspect 209-219.

Chang LJ, Sanfey AG (2013) Great expectations: neural computations underlying the use of social norms in decision-making. Soc Cogn Affect Neurosci 8:277-284.

Cheng X, Zheng L, Li L, Guo X, Wang Q, Lord A, Hu Z, Yang G (2015) Power to Punish Norm Violations Affects the Neural Processes of Fairness-Related Decision Making. Front Behav Neurosci 9:344.

Civai C, Crescentini C, Rustichini A, Rumiati RI (2012) Equality versus self-interest in the brain: differential roles of anterior insula and medial prefrontal cortex. NeuroImage 62:102-112.

Corradi-Dell'Acqua C, Civai C, Rumiati RI, Fink GR (2013) Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. Soc Cogn Affect Neurosci 8:424-431.

Feng C, Luo YJ, Krueger F (2015) Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis. Hum Brain Mapp 36:591-602.

Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. J Econ Behav Organ 3:367-388.

Gabay AS, Radua J, Kempton MJ, Mehta MA (2014) The Ultimatum Game and the brain: a meta-analysis of neuroimaging studies. Neurosci Biobehav Rev 47:549-558.

Gong X, Huang Y, Wang Y, Luo Y (2011) Revision of the Chinese facial affective picture system. Chin Mental Health J 25:40-46 (Chinses).

Guo X, Zheng L, Cheng X, Chen M, Zhu L, Li J, Chen L, Yang Z (2014) Neural responses to unfairness and fairness depend on self-contribution to the income. Soc Cogn Affect Neurosci 9:1498-1505.

Guo X, Zheng L, Zhu L, Li J, Wang Q, Dienes Z, Yang Z (2013) Increased neural responses to unfairness in a loss context. NeuroImage 77:246-253.

Hechter M, Opp KD (2001) Social norms. New York: Russell Sage Foundation.

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65-70.

Kerns JG, Cohen JD, Cho RY, Stenger VA, Carter CS (2004) Anterior cingulate conflict monitoring and adjustments in control. Science 303:1023-1026.

King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague PR (2008) The rupture and repair of cooperation in borderline personality disorder. Science 321:806-810.

Montague PR, Lohrenz T (2007) To detect and correct: Norm violations and their enforcement. Neuron 56:14-18.

ACCEPTED MANUSCRIPT

Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S (2004) The role of the medial frontal cortex in cognitive control. Science 306:443-447.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. Science 300:1755-1758.

Spitzer M, Fischbacher U, Herrnberger B, Gron G, Fehr E (2007) The neural signature of social norm compliance. Neuron 56:185-196.

Tabibnia G, Satpute AB, Lieberman MD (2008) The sunny side of fairness - Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychol Sci 19:339-347.

Wang X, Zheng L, Cheng X, Li L, Sun L, Wang Q, Guo X (2015) Actor-recipient role affects neural responses to self in emotional situations. Front Behav Neurosci 9:83.

Wu Y, Zhou Y, van Dijk E, Leliveld MC, Zhou X (2011) Social Comparison Affects Brain Responses to Fairness in Asset Division: An ERP Study with the Ultimatum Game. Front Hum Neurosci 5:131.

Xiang T, Lohrenz T, Montague PR (2013) Computational Substrates of Norms and Their Violations during Social Exchange. J Neurosci 33:1099-1108.

Zheng L, Guo X, Zhu L, Li J, Chen L, Dienes Z (2015) Whether others were treated equally affects neural responses to unfairness in the Ultimatum Game. Soc Cogn Affect Neurosci 10:461-466.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Figure 1. Timeline of Screens in a Trial

At the beginning of each trial, a fixation was presented first and lasted for 1-4s. Then the next screen appeared and lasted for 6s, displaying the average amount of money the proposer offered to other responders in the upper part and the proposal the proposer made in the lower part. Next, a blank screen appeared and was jittered from 0.5s to 1.1s. Participants were told to make a decision between accepting and rejecting the offer within 3s by pressing corresponding buttons in the decision screen. As soon as they responded, a blue frame outside the selected choice would be presented and last for 1s to provide feedback of their choice.

Figure 2. Behavioral Results

(A) Rejection rates, **(B)** Fairness ratings, and **(C)** Reaction times of each kind of proposals were plotted for all three *Treatment* conditions.

(D) Reaction times for *Accept* and *Reject* trials were plotted for all three *Treatment* conditions.

More = *MoreAve*, *Equal* = *EqualAve*, *Less* = *LessAve*. Error bars indicate SEM.

Figure 3. AI Activities in the Norm-related Model

(A) AI activity in the $(Unfair - Fair)_{EqualAve}$ contrast

(B) AI activity in the *F*-contrast ($Unfair_{MoreAve}$, $Unfair_{EqualAve}$ and $Unfair_{LessAve}$)

(C) AI activity in the $(Both - Single)_{Violated}$ contrast

(D) AI activity in the $(Single - Non)_{Violated}$ contrast

More = *MoreAve*, *Equal* = *EqualAve*, *Less* = *LessAve*. All the activations survived the voxel-level threshold of Family-wise error (FWE) corrected $p < 0.05$ with an extent threshold of 10 voxels. Error bars indicate SEM, $*p < 0.05$.

Figure 4. Brain Regions showing Main Effect and Interaction in the Response-related Model

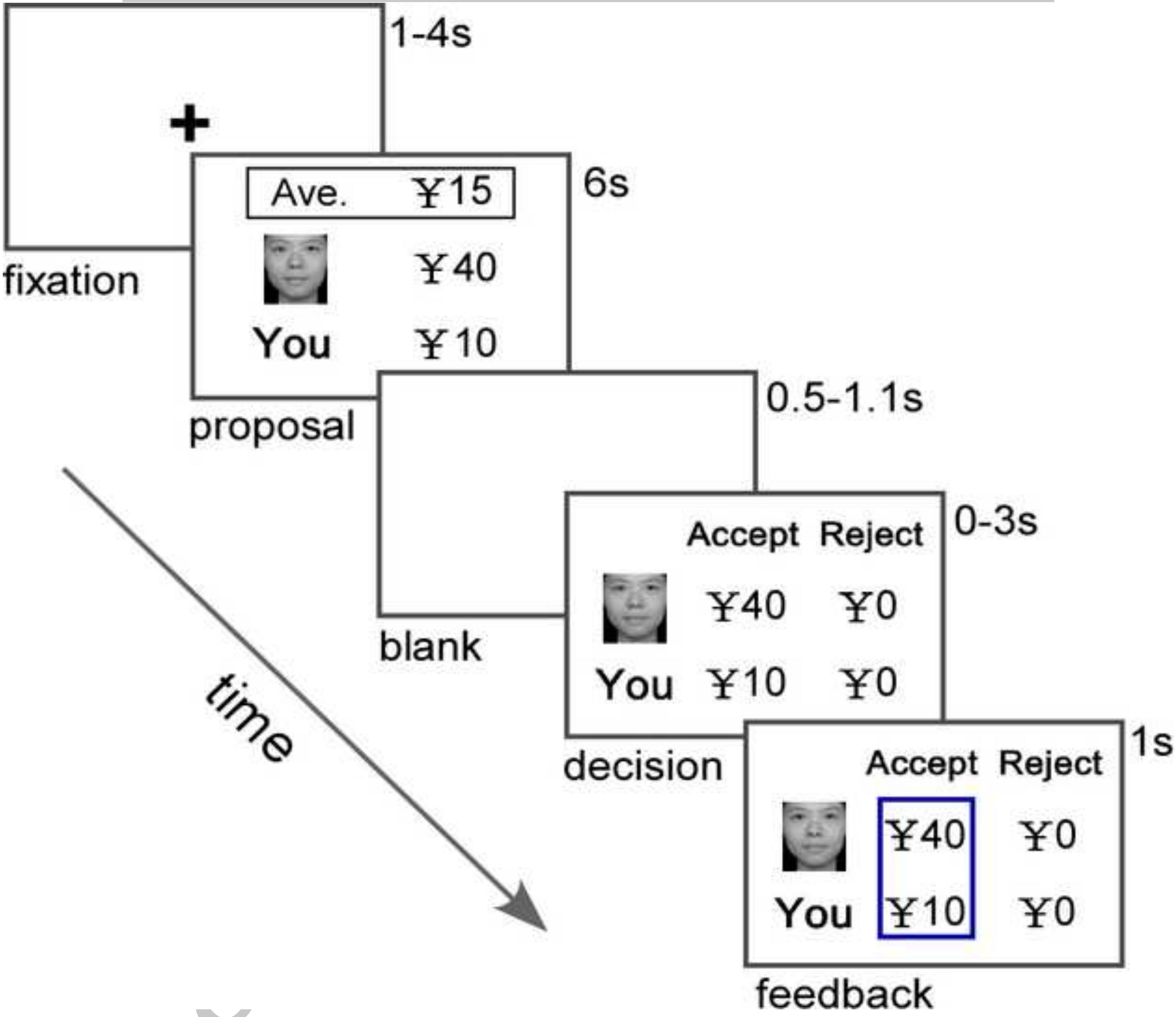
(A) AI and dmPFC/dACC activations revealed significant main effect of *Treatment.*

(B) dmPFC/dACC revealed significant interaction between *Treatment* and *Response*.

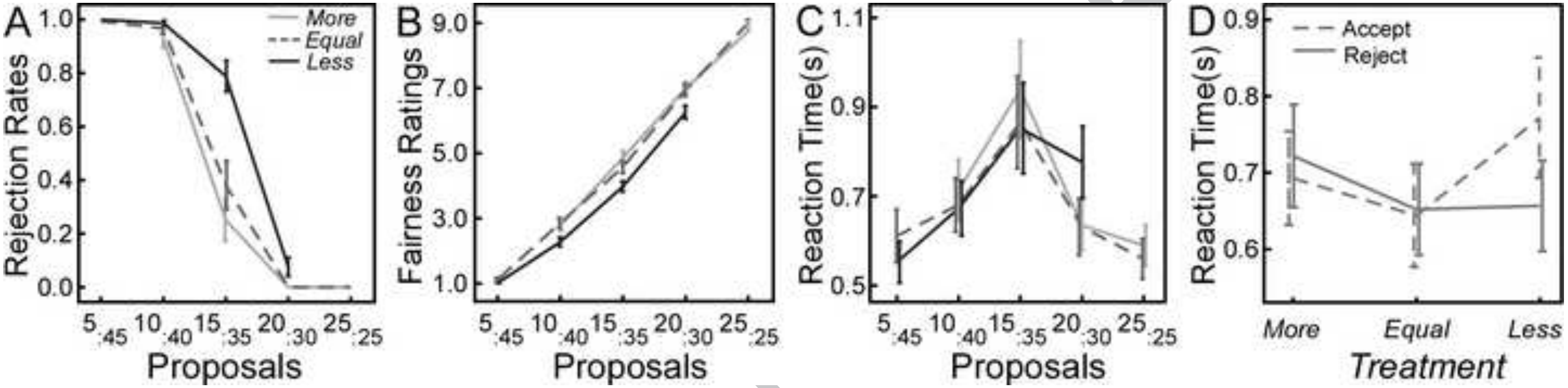
More = *MoreAve*, *Equal* = *EqualAve*, *Less* = *LessAve*. All the activations survived the voxel-level threshold of Family-wise error (FWE) corrected $p < 0.05$ with an extent threshold of 10 voxels. Error bars indicate SEM, $*p < 0.05$.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

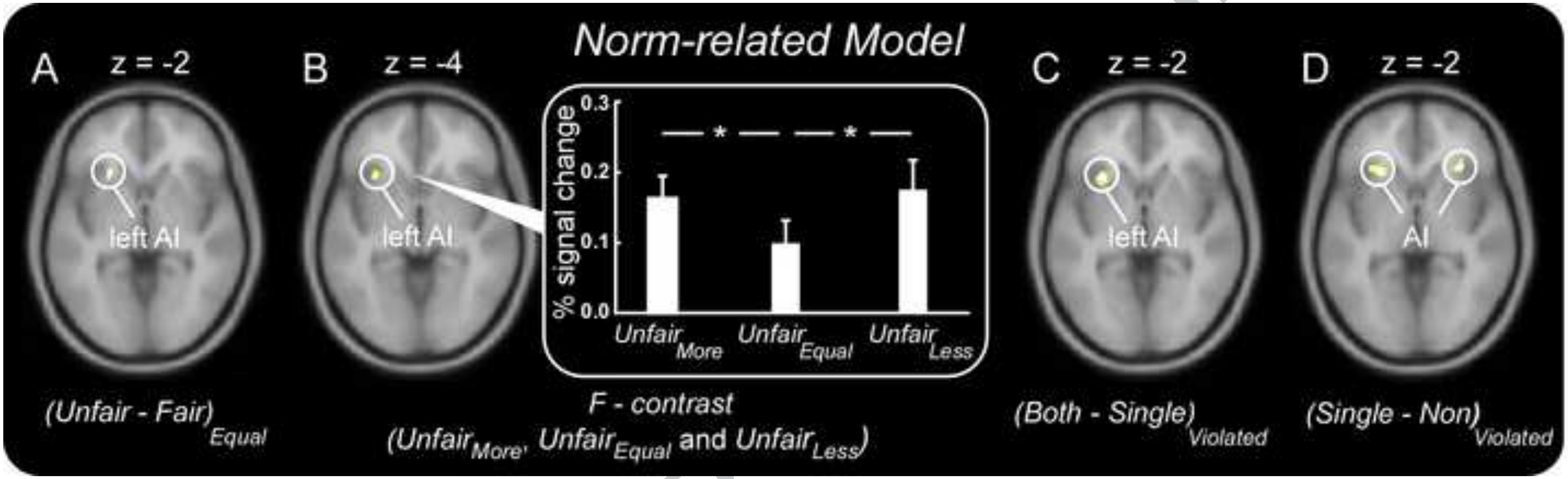


ACCEPTED MANUSCRIPT



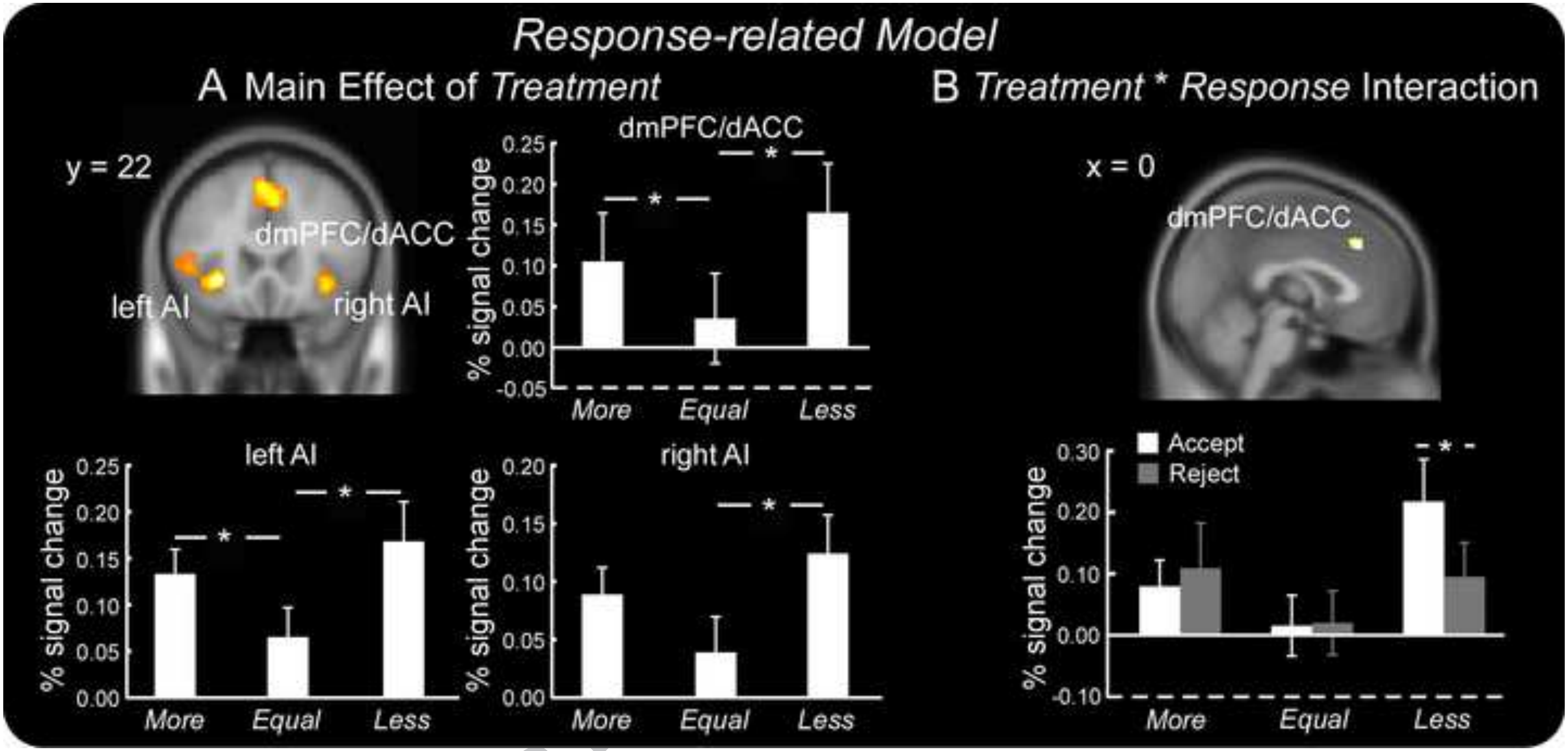
ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT



ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT



ACCEP

ACCEPTED MANUSCRIPT

Highlights

We explored how AI responded to expectation violations in a modified Ultimatum Game.

Participants were informed of both offers they received and offers others received.

AI activity increased when participants received unequal offers from proposers.

AI activity also increased when participants were offered unequally to others.

ACCEPTED MANUSCRIPT