Social Cognitive and Affective Neuroscience, 2019, 539–548

doi: 10.1093/scan/nsz031
Advance Access Publication Date: 29 April 2019
Original article

Born for fairness: evidence of genetic contribution to a neural basis of fairness intuition

Yun Wang,[1,2,5] Dang Zheng,[1,2] Jie Chen,[2,3] Li-Lin Rao,[1,2] Shu Li,[1,2,4] and Yuan Zhou[1,2,4]

[1]CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China, [2]Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China, [3]CAS Key Laboratory of Mental Health, Institute of Psychology, Beijing 100101, China, [4]Magnetic Resonance Imaging Research Center, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China, and [5]The National Clinical Research Center for Mental Disorders & Beijing Key Laboratory of Mental Disorders, Beijing Anding Hospital, Capital Medical University, Beijing 100088, China.

Correspondence should be addressed to Yuan Zhou, CAS Key Laboratory of Behavioral Science Institute of Psychology, Chinese Academy of Sciences, No. 16 Lincui Road, Chaoyang District, Beijing 100101, PR China. Email: zhouyuan@psych.ac.cn.

Abstract

Human beings often curb self-interest to develop and enforce social norms, such as fairness, as exemplified in the ultimatum game (UG). Inspired by the dual-system account for the responder's choice during the UG, we investigated whether the neural basis of psychological process induced by fairness is under genetic control using a twin fMRI study (62 monozygotic, 48 dizygotic; mean age: 19.32 ± 1.38 years). We found a moderate genetic contribution to the rejection rate of unfair proposals (24%–35%), independent of stake size or proposer type, during the UG. Using a voxel-level analysis, we found that genetic factors moderately contributed to unfairness-evoked activation in the bilateral anterior insula (AI), regions representing the intuition of fairness norm violations (mean heritability: left 37%, right 40%). No genetic contributions were found in regions related to deliberate, controlled processes in the UG. This study provides the first evidence that evoked brain activity by unfairness in the bilateral AI is influenced by genes and sheds light on the genetic basis of brain processes underlying costly punishment.

Key words: ultimatum game; twin study; heritability; fMRI; social norm

Introduction

Human beings often curb self-interest to develop and enforce social norms, such as fairness, which is considered essential for the evolution of cooperation in human beings (Fehr and Schmidt, 1999; Camerer, 2003). A canonical example is the ultimatum game (UG), in which one player (proposer) proposes a division of a sum of money between himself/herself and a second player (responder), who either accepts or rejects it (Güth *et al.*, 1982). If the responder accepts the proposal, the suggested split is realized. If the responder rejects the offer, neither of the two receives anything. While the responders face an unfair proposal, they have to trade off between the self-interest motive and a fairness preference (Knoch *et al.*, 2006). Rejection of unfair proposals means that the responders succeed in curbing their self-interest motive, i.e. maximizing their economic gain, to pursue fairness.

Received: 11 May 2018; **Revised:** 7 February 2019; **Accepted:** 21 April 2019

Rejection of unfair proposals is observed across countries and ethnicities (Camerer, 2003; Henrich *et al.*, 2005), which has triggered the discussion on the biological origin of this norm-enforcement behavior. A behavioral genetics study on Swedish twins suggested that this decision is controlled by genes, as >40% of the variation in the rejection behavior of responders was explained by additive genetic effects (Wallace *et al.*, 2007). However, this behavioral genetics study cannot answer which psychological process subserving this decision is under genetic control.

Researchers have tried to understand the psychological processes underlying the social preferences of the responder in fairness-related norm enforcement. Early models have postulated the significant roles of inequality aversion (Fehr and Schmidt, 1999) and intention inference (Rabin, 1993; Blount, 1995). In recent decades, these initial theoretical models have been extensively elaborated due to interdisciplinary studies in the fields of psychology, economics and neuroscience (Sanfey, 2007; Rilling and Sanfey, 2011). Recent studies have attempted to understand fairness-related norm enforcement in response to norm violations from the perspective of dual-system theories (Sanfey *et al.*, 2006; Sanfey and Chang, 2008; Buckholtz and Marois, 2012; Feng *et al.*, 2015), which have received extensive theoretical consideration in the field of cognition (Evans, 2003; Lieberman, 2007) and judgment and decision-making (Evans, 2008; Sanfey and Chang, 2008). In dual-system theories, System 1, which is automatic and heuristic-based and quickly proposes intuitive answers to problems as they arise, includes the anterior insula (AI), the dorsal anterior cingulate cortex (ACC) and the ventromedial prefrontal cortex; and System 2, which corresponds closely to controlled processes, monitors the quality of the answer provided by System 1 and sometimes corrects or overrides these judgments, includes the ventral ACC, the lateral prefrontal cortex (PFC), lateral parietal cortex and dorsomedial PFC (Satpute and Lieberman, 2006; Lieberman, 2007). In terms of the decision-making of the responder in the UG, researchers consistently observed that the regions relevant to the dual-system theories are more activated when the participants face unfair proposals than when they face fair proposals and thus proposed that System 1 represents the psychological components involved in rapidly evaluating violations of the fairness norm; and System 2 is involved in integrating both self-interest and the fairness norm to regulate the intuitive system to permit more flexible decision-making (Sanfey *et al.*, 2006; Feng *et al.*, 2015). However, this appealing proposal omits another possible candidate intuition possibly implicated in System 1, i.e. monetary self-interest, because most of the previous studies cannot distinguish between fairness and monetary incentives. In other words, in these studies, an unfair offer is one with lower monetary incentives and a fair offer is one with higher monetary incentives (Sanfey *et al.*, 2003; Chang and Sanfey, 2011; Corradi-Dell'Acqua *et al.*, 2012; Xiang *et al.*, 2013). Therefore, the rejection of unfair offers may result from two possibilities: it is possible that the fairness intuition drives the participant to make a judgment whether the offer violates a fairness norm and thus generate an impulse to reject unfair offers; it is also possible that a monetary self-interest intuition generates the same decision by simply judging whether the offered amount is lower than expected or absolute value of the reward. Similarly, the stronger brain activity seen when facing unfair offers than that when facing fair offers in this situation cannot exclude the possibility that the observed activation is due to unexpected small monetary incentives. To exclude the influence of monetary self-interest on the social decision-making of the responder, several previous studies applied a revised UG paradigm, in which the same amount of monetary incentive may be fair or unfair (Tabibnia *et al.*, 2008; Zhou *et al.*, 2014).

Armed with this revised paradigm, in this study, we aim to investigate whether the neural basis of psychological process induced by fairness during the UG is under genetic control. Although the dual-system theories assume that the processes in System 2 may be heritable based on its close relationship with genetically determined general intelligence and working memory (Evans, 2008), few empirical evidence support this hypothesis (Sanderson *et al.*, 2009). On the other hand, the processes in System 1 are considered as universal (Evans, 2008), as we observed an unfair proposal is always perceived as unfair, even though the extent of unfairness may be modulated by experimental factors, such as proposer type or stake size (Zhou *et al.*, 2014). However, the universality cannot exclude the possibility that these processes are heritable.

To investigate the possibility of such a genetic basis, we conducted a twin fMRI study, which is a powerful tool in establishing the heritability of phenotype (Martin *et al.*, 1978; Neale and Cardon, 2013). First, we estimated the genetic contribution to responders' behavior in a UG by orthogonally manipulating fairness and stake size from human or computer partners to examine whether the norm enforcement indicated by rejection of unfairness is genetic independent of experimental factors, such as stake size and proposer type. Then, we investigated in which region(s) the individual variation in brain activation induced by fairness during the UG is attributable to genetic or environmental influences by using a voxel-wise genetic modeling analysis. This voxel-wise analysis makes it possible to search the whole brain and identify region-specific effects and answer the question of the neural basis of which psychological process is heritable.

Materials and methods

Participant

A total of 110 same-sex twin pairs (sex: 50.91% male; age: M = 19.32, s.d. = 1.38 years) sampled from the Beijing Twin Study (BeTwiSt) participated in this study, among which 62 pairs were monozygotic (MZ) and the other 48 pairs were dizygotic (DZ). For all twin pairs who participated in our study, zygosity was assigned by DNA testing, with a classification accuracy of nearly 100% (Chen *et al.*, 2010).

All participants were in good health, with no previous history of psychiatric or neurological disease based on their self-reports. Written informed consent was obtained following a detailed explanation of the study. The participants were given a financial reward at the end of the study. The study was approved by the institutional review board of the Institute of Psychology, Chinese Academy of Sciences, and the institutional review board of the Beijing MRI Center for Brain Research.

Procedure and experimental design

Before scanning, the participants received instructions explaining the rules of the game and were required to answer a series of questions after reading the instructions to verify their comprehension. During scanning, the participants acted as responders to play a one-shot game with a different proposal for each trial. After completing the UG task, the participants rated the fairness of all offers presented in the UG task on a Likert scale of 1 (very unfair) to 7 (very fair). To increase the degree of involvement in

Table 1. Types of offers

	Fair (50%)	Unfair (20%)
High stake size (¥)	400/450/500/550/600 out of 800/900/1000/1100/1200	400/450/500/550/600 out of 2000/2250/2500/2750/3000
Low stake size (¥)	4/4.5/5/5.5/6 out of 8/9/10/11/12	4/4.5/5/5.5/6 out of 20/22.5/25/27.5/30

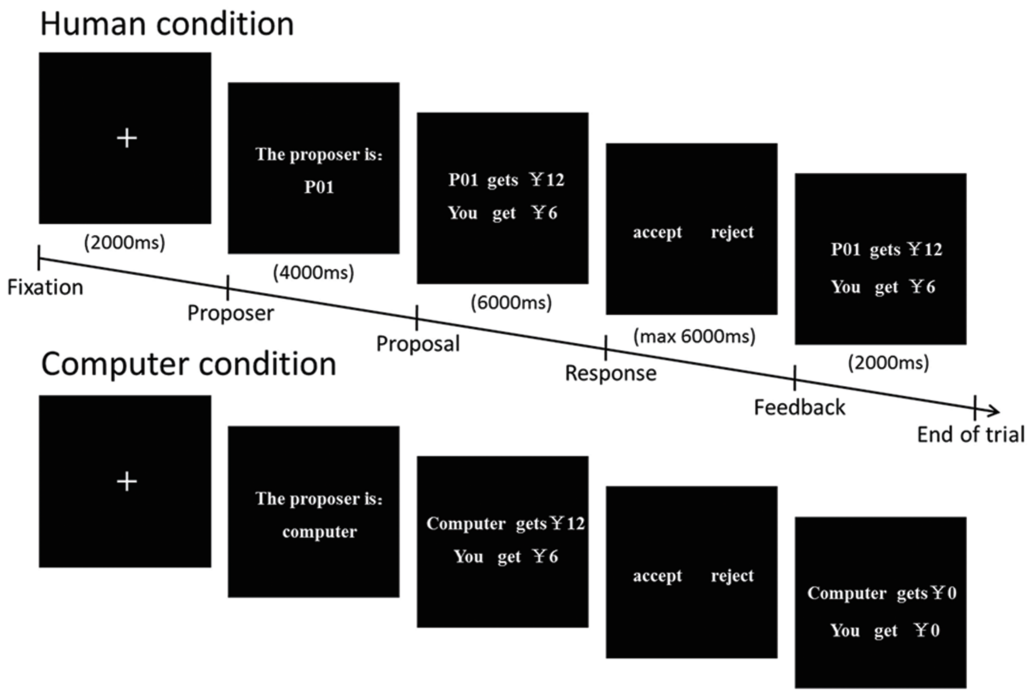

Fig. 1. Timeline for a single round of the UG.

this task, the participants also made proposals with different stake sizes as proposers after the experiment and were told that their proposals would be used in a subsequent study.

To differentiate which psychological process induced by fairness during the UG is under genetic control, we designed a repeated one-shot UG task, which was similar to our previous study (Zhou *et al.*, 2014). Since responders often reject unfair proposals and accept fair proposals, we set two 'fairness' categories: 50% of the stake (fair) and 20% of the stake (unfair) to differentiate the genetic contribution to the responder's normative decisions. In addition, we set two factors to modulate the responder's decision. One was the stake size, which was set orthogonally to fairness by varying both the proposal amount and the stake size across the rounds. The proposal amount to the responder (i.e. monetary self-interest) was fixed for fair and unfair proposals at the same level of stake size. The other was proposer type. The participants were offered proposals from real persons, who participated in the experiment and submitted their proposals, or from computer partners, which generated the proposals randomly. In reality, all the offers were pre-set by the experimenter. The offers from the computer partners were identical to those from the human partners, and the computer condition was similar to the human condition in terms of fairness and self-interest, except for the fact that there was no potential for social interaction in the computer condition. There were four combinations of offer size and fairness in each proposer condition, and five rounds were run for each combination. The types of offers can be seen in Table 1. Therefore, the responders played 40 rounds, 20 of which were supposedly from a game with human partners and 20 with computer partners (Figure 1). The proposals from human and computer partners were presented randomly. To encourage participants to make real decisions, it was emphasized that in addition to a fixed amount for participation, they would be paid according to their choices in the game.

fMRI data acquisition

The fMRI data were acquired from the Beijing MRI Center for Brain Research. MR images sensitized to changes in blood oxygen level dependent (BOLD) signal levels were obtained by an echo planar imaging sequence on a 3.0-Tesla Siemens MR scanner (repetition time = 2000 ms; echo time = 30 ms; flip angle = 90 degrees, matrix = 64×64; field of view = 220×220 mm^2; slice thickness = 3 mm; slice gap = 1 mm). Each brain volume was composed of 32 axial slices. The scanning duration depended on the participant's response and ranged from 324 TR to 347 TR (average scanning duration = 327 TR). Stimuli were presented with E-prime software (Psychology Software Tools, Pittsburgh, PA, USA) on a personal computer, back-projected onto a screen using a liquid crystal display projector and viewed by the participants through a mirror mounted on the MRI head coil. The scanner was triggered by a signal generated by E-prime stimulus

presentation software to synchronize each volume acquisition with the onset of a visual stimulus.

fMRI data processing

Image preprocessing was performed using statistical parametric mapping (SPM8, Wellcome Department, London, UK). The preprocessing included slice time correction, realignment, normalization, resampling to $3 \times 3 \times 3$ mm^3 and smoothing using an 8 mm full-width-at-half-maximum Gaussian kernel. Subjects with head motion >3 mm in translation or 3 degrees in rotation were labeled and repaired by using the ArtRepair toolbox (http://cibsr.stanford.edu/tools/human-brain-project/artrepair-software.html) (Mazaika *et al.*, 2009). Images with artifacts were repaired, and the quality checks were calculated and detected. Subjects with improved data quality after repair were re-incorporated into the analysis, while subjects who could not be corrected by ArtRepair were excluded from the analysis. Finally, 193 subjects were included in the fMRI analysis, and there were 85 twin pairs (49 MZ and 36 DZ) among them.

A general linear model (GLM) with a 2 (fairness) × 2 (proposer type) × 2 (stake size) factorial design matrix was constructed to detect the brain activation of each participant during the proposal epochs. Specifically, a GLM was defined for each participant. These models included eight regressors that modeled the BOLD response to the 6 s proposal epoch: fair proposal from a human partner, unfair proposal from a human partner, fair proposal from a computer partner and unfair proposal from a computer partner for each of the high and low stake sizes. Additionally, six motion parameters obtained by realignment were used as nuisance variables. Each regressor was convolved with a canonical hemodynamic response function. High-pass filtering (cutoff frequency = 128 s) was used to remove low-frequency noise. The resulting GLM was corrected for temporal autocorrelations using a first-order autoregressive model. First-level contrasts were performed for each experimental condition of the factorial design described above. To account for the dependency between twins in the same pair, we emulated a hierarchical linear model (HLM) using the standard summary statistic approach before conducting a second-level random-effect analysis. Specifically, we first averaged the first-level contrast images for each twin pair and then used the averages as data for the second-level random-effect analysis to detect group effects. In this study, we were particularly interested in fairness-related brain activity, including activation evoked by unfair proposals compared to fair proposals (unfair > fair *t* contrast) and activation evoked by fair proposals compared to unfair proposals (fair > unfair *t* contrast), regardless of zygosity. For the whole brain, significant activations were required to exceed a height threshold of $P < 0.05$ after family-wise error (FWE) corrected for multiple comparisons and cluster-size threshold of 10 voxels.

Genetic modeling

By comparing the resemblance of MZ and DZ twin pairs on observed trait(s), we estimated additive genetic (A), common (shared) environmental (C) and non-shared environmental (E) contributions to variance within a trait (Plomin *et al.*, 2013). Correlations between additive genetic factors are fixed at 1 for MZ twin pairs, as they share 100% of their genes, and at 0.5 for DZ pairs as they share, on average, 50% of their genes. In the case that twins are reared together, the greater resemblance between MZ twins than that between DZ twins indicates that the trait is heritable. The proportion of trait variance explained by additive genetic effects is referred to as heritability. By definition, common environmental factors are those factors in the environment that make twins growing up in the same family similar to each other. For common environmental factors, correlations between co-twins are fixed at 1 for both MZ and DZ pairs, based on the rigorous and frequent testing that has supported the assumption that environments for MZ and DZ twins are comparable. Non-shared environmental factors are those factors that make twins less similar to each other, including environmental factors unique to each individual and measurement error. They are left uncorrelated in twins.

Genetic modeling of responder's normative decision. To estimate genetic and environmental effects on the responder's normative decisions, we used the rejection rate as the dependent variable to conduct univariate genetic modeling implemented in the OpenMx package for R (http://openmx.psyc.virginia.edu). First, we calculated the intraclass correlation coefficient (ICC) for the MZ and DZ twins separately. If ICC_{MZ} was greater than ICC_{DZ}, this suggested that MZ twins resembled each other more than DZ twins. We then used univariate models to partition the variance of this measure into genetic (A) and environmental (C and E) effects. We examined the full ACE model first. Submodels (AE, CE and E) nested within the full model were then tested by systematically removing one or two components of the variance. We used the change in chi-square (χ^2) and the Bayesian information criterion (BIC) as model fit indices (Raftery, 1995). A lower BIC value indicates better fit. Comparing the full model and a sub-model, a significant χ^2 difference suggested that the nested model fit significantly worse than the full model and the full model should be chosen; otherwise, the nested model with fewer parameters should be considered in terms of parsimony (Bollen, 1989; Kline, 1998).

Genetic modeling of brain activity. Using a similar procedure, we conducted a voxel-wise genetic modeling of the brain activity. As to the fairness-related brain activity, we restricted the genetic modeling analyses to voxels, which were specified by the group analysis and showed greater intraclass correlations for MZ twins (ICC_{MZ}) than DZ twins (ICC_{DZ}). We fitted univariate genetic modeling voxel by voxel to estimate the contributions of A, C and E to explain the variance in fairness-related brain activation, and then submodels (AE, CE, and E) nested within the full model were tested by systematically removing one or two components of variance. For almost all the voxels, the best-fitting model was AE (see result). Then, we assessed the genetic influence (i.e. the A component) in terms of the difference in log-likelihood after it was removed (i.e. comparing the AE model and the E model), using the goodness-of-fit χ^2 statistic. This likelihood enabled us to construct posterior probability maps (PPMs) to identify regions showing a genetic effect with ≥95% posterior confidence (Friston and Penny, 2003). The construction of PPMs enables Bayesian inferences about regionally specified effects in neuroimaging. The PPMs report the posterior probability or confidence that an effect exceeds some specified confidence level, given the data. In contrast to classical inference, which is based on rejecting the null hypothesis, PPMs report the posterior probability that an effect is present (with a small probability that it is not). This means there are no declaration of a 'significant' effect, no false-positive rate and no multiple-comparisons problem. This application of PPMs in twin fMRI analysis has been reported in a previous study (Rao *et al.*, 2018).

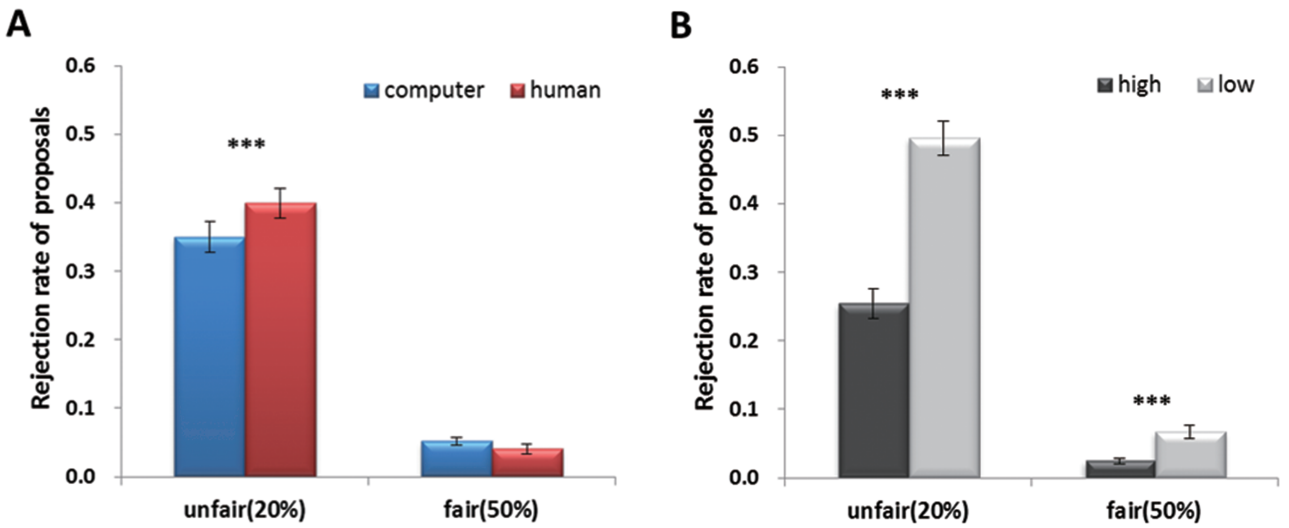

Fig. 2. Mean rejection rates as a function of proposal fairness for different proposer types (A) and stake sizes (B). Error bars represent the SE of the difference of the means.

We also conducted the same procedure for brain activity induced by proposer type, stake size or interaction effects.

Results

Genetic contribution to responder's decision

To determine whether the rejection rate was influenced by experimental factors, we first investigated the main effects of fairness, proposer type and stake size and the interaction effects between these factors on rejection rate using repeated-measures analysis of variance (ANOVA). Significant main effects of fairness [$F(1,219) = 237.11$, $P < 0.001$, partial $\eta^2 = 0.52$], proposer type [$F(1,219) = 6.66$, $P = 0.011$, partial $\eta^2 = 0.03$] and stake size [$F(1,219) = 124.89$, $P < 0.001$, partial $\eta^2 = 0.36$] were found. These main effects separately indicated that unfair proposals (M = 0.37, s.d. = 0.31) were more often rejected than fair ones (M = 0.05, s.d. = 0.09), proposals from humans (M = 0.22, s.d. = 0.17) were more often rejected than those from computers (M = 0.20, s.d. = 0.18), and proposals with a low stake size (M = 0.28, s.d. = 0.21) were more often rejected than those with a high stake size (M = 0.14, s.d. = 0.17). In addition, the interaction between fairness and proposer type was significant [$F(1,219) = 16.83$, $P < 0.001$, partial $\eta^2 = 0.07$]. A post hoc pairwise least significant difference (LSD) test indicated that the rejection rates for proposals from human partners were significantly higher than those for proposals from computer partners when the proposals were unfair ($P < 0.001$, Figure 2A). The interaction between fairness and stake size was also significant [$F(1,219) = 79.44$, $P < 0.001$, partial $\eta^2 = 0.27$]. A post hoc pairwise LSD test indicated that the rejection rate for proposals with a low stake size was significantly higher than that for proposals with a high stake size in both the fair and unfair proposal condition (Ps < 0.001, Figure 2B). No interaction between fairness, proposer type and stake size was found.

Accounting for the dependency between twins in the same pair, we used an HLM to validate the abovementioned behavioral findings. HLM is an ideal method for analyzing twin data because it allows for nested data analysis, accounting for the correlated nature of twin data (Lynch *et al.*, 2006; Keuler *et al.*, 2011; Lydecker *et al.*, 2012). We applied a two-level HLM to assess the effects of fairness, stake size, proposer type and the two-way and three-way interactions on rejection rate. Individual twins were the first-level unit nested inside the 'family' variable shared by co-twins. Regression equations were computed to predict the rejection rate using the above independent variables as dichotomous predictors. Regression coefficients and *p* values were reported for each of the predictors in Table 2. Compatible with the ANOVA, we found that fairness, stake size, proposer type, interaction between fairness and stake size, interaction between fairness and proposer type and the three-way interaction were significant predictors of rejection rate.

Because the rejection rate of proposals was modulated by proposer type or stake size, we separately analyzed the genetic contribution to rejection rate under each condition. In general, for the rejection rate of unfair proposals, the MZ twin correlation was significantly higher than the DZ correlation, whether they were from a human or computer partner or with a large or small stake size (Table 3), suggesting that genes make a substantial contribution to the individual differences in terms of costly punishment. By conducting univariate model-fitting analyses for the conditions including unfair proposals, we found the AE model was the best model to partition the phenotypic variance. The AE model attributed 24%–35% of individual difference in the rejection rate of unfair proposals due to genetic influences and the other 65%–76% to non-shared environmental influences (Table 4), suggesting a moderate heritability for costly punishment of responder during the UG. More importantly, a genetic effect on the rejection rate of unfair proposals existed, whether from a human or computer partner or with a low or high stake size, suggesting that the norm enforcement indicated by rejection of unfairness is genetic independent of modulator factors (proposer type and stake size).

For the rejection rate of fair proposals, unlike that for unfair proposals, the MZ twin correlations were not significant and were lower than the DZ correlations, indicating an absence of genetic influence and the dominant role of environmental factors in determining the individual difference in the rejection rate of fair proposals (Table 3). No further genetic modeling analyses were done for this case. Similarly, we found that environmental factors rather than genetic factors contributed to the modulation effects of proposer type or stake size on rejection rate (Table 3).

Table 2. Results from the HLM examining the influence of fairness, stake size, proposer type and the interactions between them on rejection rate

Parameter	B	SE	T	*df*	P
Fairness	0.164	0.012	13.895	109	<0.001
Stake size	0.071	0.007	10.903	109	<0.001
Proposer type	0.009	0.004	2.504	109	0.014
Fairness∗stake size	0.050	0.006	8.523	109	<0.001
Fairness∗proposer type	0.015	0.004	3.892	109	<0.001
Stake size∗proposer type	0.002	0.003	0.782	109	0.436
Fairness∗stake size∗proposer type	0.019	0.003	6.217	109	<0.001

Note: B, unstandardized regression coefficient; SE, standard error; *df*, degree of freedom.

Table 3. Mean (s.d.) rejection rate under each condition and twin ICCs (95% confidence intervals)

Rejection rate	Mean (s.d.)	Twin correlation		Fisher's Z test
		ICC_{MZ}	ICC_{DZ}	
unfair_human	0.40(0.32)	0.51**(0.19∼0.71)	0.15(−0.52∼0.52)	2.08*
unfair_computer	0.35(0.33)	0.56***(0.26∼0.73)	−0.05(−0.88∼0.41)	3.45***
unfair_high	0.25(0.32)	0.45**(0.09∼0.67)	−0.18(−1.11∼0.34)	3.37***
unfair_low	0.50(0.38)	0.52**(0.20∼0.71)	0.31(−0.24∼0.61)	1.29
fair_human	0.04(0.10)	−0.02(−0.70∼0.38)	0.66***(0.39∼0.81)	−4.11***
fair_computer	0.05(0.09)	0.26(−0.23∼0.55)	0.36(−0.15∼0.64)	−0.56
fair_high	0.03(0.06)	0.28(−0.19∼0.57)	0.74***(0.53∼0.85)	−3.35***
fair_low	0.07(0.14)	0.08(−0.53∼0.45)	0.21(−0.42∼0.56)	−0.67
fairness∗proposer type	0.06(0.22)	0.02(−0.62∼0.41)	0.10(−0.61∼0.49)	−0.41
fairness∗stake size	−0.20(0.33)	0.25(−0.25∼0.55)	0.11(−0.59∼0.50)	0.73

Note: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$

Table 4. Univariate genetic modeling for rejection rate under each condition

Rejection rate	Model	-2LL	*df*	BIC	Change from full model			A	C	E
					$\triangle\chi^2$	$\triangle df$	P			
unfair_human	ACE	112.49	216	−902.82				0.34 (0.00–0.53)	0.00 (0.00–0.35)	0.66 (0.47–0.89)
	AE	112.49	217	−907.52	0	1	1	0.34 (0.11–0.53)		0.66 (0.47–0.89)
unfair_computer	ACE	120.87	216	−894.43				0.34 (0.00–0.54)	0.00 (0.00–0.26)	0.66 (0.46–0.90)
	AE	120.87	217	−899.14	0	1	1	0.34 (0.10–0.54)		0.66 (0.46–0.90)
unfair_high	ACE	125.24	216	−890.06				0.24 (0.00–0.47)	0.00 (0.00–0.24)	0.76 (0.53–1.00)
	AE	125.24	217	−894.76	0	1	1	0.24 (0.00–0.47)		0.76 (0.53–1.00)
unfair_low	ACE	184.47	216	−830.83				0.34 (0.00–0.53)	0.01 (0.00–0.42)	0.65 (0.47–0.88)
	AE	184.47	217	−835.53	0	1	0.97	0.35 (0.14–0.53)		0.65 (0.47–0.86)

Note: The full ACE model and the best-fitting model are presented for each condition. −2LL, twice the negative log-likelihood; $\triangle\chi^2$, change in chi-square; $\triangle df$, change in degrees of freedom; A, proportion of variance due to additive genetic effects; C, proportion of variance due to shared environmental effects; E, proportion of variance due to non-shared environmental effects. The 95% confidence intervals are in parentheses.

Together, these findings suggest that the costly punishment for unfairness has an innate mechanism, which is independent of some experimental factors (such as proposer type and stake size).

Genetic contribution to brain activation

To investigate the genetic contributions of the neural basis of psychological processes induced by fairness during the UG, we first identified the brain regions whose activities were modulated by fairness. Specifically, we found that the bilateral insular cortices, striatum, medial PFC extending to the anterior cingulate cortex (ACC), lateral PFC, inferior parietal cortex, superior parietal cortex and middle occipital gyrus showed greater activation in response to unfair proposals than to fair proposals (FWE corrected $P < 0.05$; voxels, >10; Figure 3B).

Intraclass correlations for unfairness-evoked brain activation are shown in Figure 4. Overall, the MZ correlations were greater than the DZ correlations, suggesting that the individual variation in unfair-evoked activation is genetically influenced. Voxel-wised genetic modeling further showed that an AE model better fit the data for most voxels (92%). In the AE model, genetic contributions to the brain activity evoked by unfairness were found in the left (mean heritability = 0.37) and right AI (mean heritability = 0.40) and the right middle occipital gyrus (mean heritability = 0.42), with ≥95% posterior confidence to support a genetic effect (Figure 5).

We also found that the bilateral middle temporal gyrus, the bilateral inferior parietal lobule, the medial prefrontal cortex and the bilateral precuneus showed greater activation in response to fair proposals than unfair proposals (FWE corrected $P < 0.05$; voxels, >10; Figure 3A). In addition, we found the main

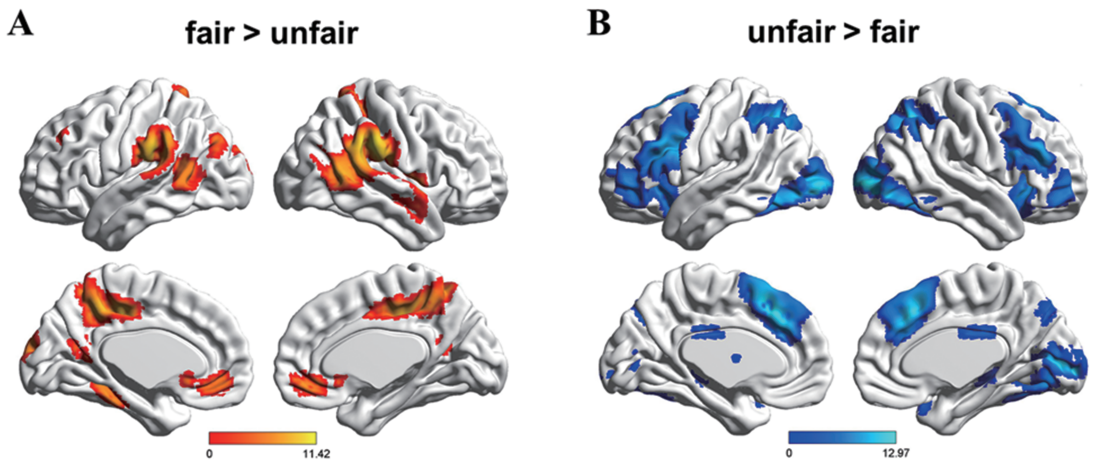

Fig. 3. Brain activations influenced by fairness at proposal presentation. (A) Maps of the *t* statistics for the contrast [fair > unfair] showing activation of the bilateral middle temporal gyrus, the bilateral inferior parietal lobule, the medial frontal gyrus and the bilateral precuneus. (B) Maps of the *t* statistic for the contrast [unfair > fair] showing activation of the bilateral insular cortices, striatum, medial prefrontal cortex extending to ACC, lateral PFC, inferior parietal cortex, superior parietal cortex and middle occipital gyrus.

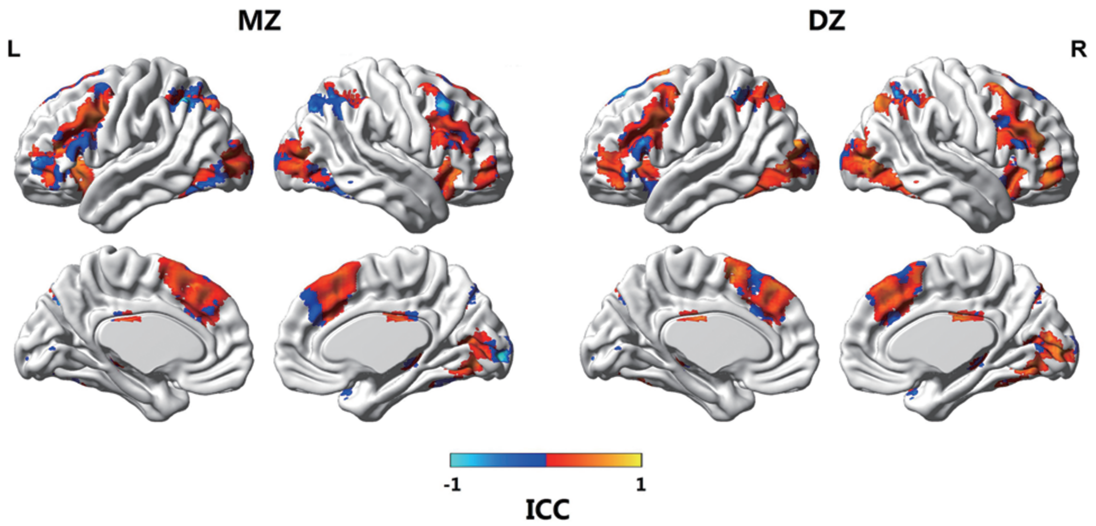

Fig. 4. ICCs for unfairness-evoked brain activation in MZ and DZ twins.

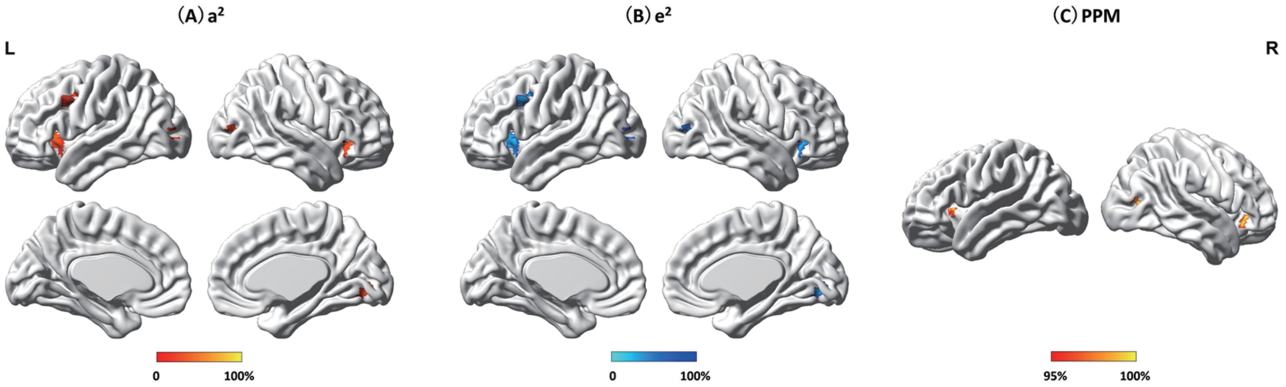

Fig. 5. Variance component estimates for unfairness-evoked brain activation. (A and B) Percentages of variance explained by genetic (a^2) and unique environmental factors (e^2) within a mask in which ICC_{MZ} was larger than ICC_{DZ}. (C) PPMs for a^2, indicating which genetic estimates were significant at the ≥95% confidence level.

effects of proposer type and of the stake size and their interaction effect (for details, please see Supplementary Figures S1, S2 and S3). No interaction effect between fairness and stake size or a three-way interaction effect was found with this strict threshold. Using a voxel-wise genetic modeling analysis, we found no strong evidence to support a genetic contribution either in regions showing greater activity when facing a fair proposal than an unfair proposal or in regions showing the main effect of proposer type and of the stake size and their interaction effect.

Discussion

This study aimed to investigate genetic contributions to the neural basis of psychological processes induced by unfairness during the UG. We found that the rejection decision for unfair proposals was heritable independent of stake size or proposer type. Furthermore, we found that genetic contributions to the brain activity evoked by unfair compared to fair proposals during the UG located in the bilateral anterior insular cortices. These findings suggest that the psychological process supported by the anterior insular cortex during the UG was heritable.

We implemented identical monetary payoff for fair and unfair proposals in the UG task. This design makes it possible to separately examine the impact of fairness and monetary self-interest on the decisions of the responders. We found that fairness *per se* can affect the decision-making of the responder in the UG after excluding the influence of monetary incentives. In addition, the rejection decision for unfair proposals was heritable in our Han ethnic twins, which is consistent with the findings in a Swedish population (Wallace *et al.*, 2007). This repeatable observation across ethnicities indicates that genes account for the part of inter-individual differences involved in deciding whether to punish others by costing themselves. Particularly, we found that the genetic contribution to the rejection of unfair proposals was independent of stake size or proposer type, suggesting that this influence stably existed in different social contexts, such as the proposal being from a real person or a computer and with different stake sizes (high or low).

The main concern of the current research is to investigate the neural basis by which psychological process is under genetic control when the participants face unfair proposals during the UG. When we compared the brain activity evoked by unfair proposals and that evoked by fair proposals during the UG, we found that the bilateral anterior insular cortices, lateral PFC, lateral parietal cortex and dorsal ACC showed stronger activity, consistent with previous studies (Sanfey, 2007; Feng *et al.*, 2015). Based on dual-system theories (Satpute and Lieberman, 2006; Lieberman, 2007), in our UG task, the regions involved in System 1 included the bilateral AI and dorsal ACC and the regions in System 2 included the lateral PFC and lateral parietal cortex. Furthermore, we examined the genetic contribution to these unfairness-evoked brain activities to uncover the neural basis of this fairness normative decision. Among these regions, only the activities of the bilateral AI were moderately controlled by genetic factors. The AI plays a crucial role in the normative decision of responders. Previous studies emphasized on the role of AI in negative emotion and interoceptive sensation (Sanfey *et al.*, 2003; Harlé and Sanfey, 2007; Grecucci *et al.*, 2012; Harlé *et al.*, 2012); however, recent evidence suggests its role in cognitive heuristics to detect norm violations (Civai *et al.*, 2012; Corradi-Dell'Acqua *et al.*, 2014). As a straightforward and parsimonious account for the variety of cognitive and emotional tasks in which the AI has been found to play a role, the activation of AI in the UG can be interpreted as a signal of deviation from an expected outcome (Civai, 2013). This is consistent with a particular role of the AI in bias or error detection (Preuschoff *et al.*, 2008; d'Acremont *et al.*, 2009; Chang *et al.*, 2013; Xiang *et al.*, 2013). In the context of our UG task, there were two candidates for prediction error: the violation of fairness norm and the unexpected offer amount. Although both of these prediction errors could induce activity in the AI, we observed increased activity in the AI only when comparing unfair proposals with fair proposals, both of which were set with the same monetary incentive. Therefore, our study provides clear evidence that the AI may be activated by the bias between the externally presented proposals and the internalized social norm (fairness) of the responder and thus may reflect an intuition of fairness norm violations. This detection of norm violation happening in the initial evaluation on the proposals is one of the psychological components implicated in System 1 (Feng *et al.*, 2015).

Along these lines, our finding that genetic factors contribute to the activity of bilateral AI induced by unfairness suggests that the neural basis of intuition of fairness norm violations is under genetic control. This is compatible with development studies on fairness preference, an intuitive process. These studies found that children as young as 2 years demonstrated preferences for fairness in UGs (Li *et al.*, 2016), and a more general predisposition toward altruism even can be observed in infants (Warneken and Tomasello, 2006). The appearance of fairness preference in the first stage of life suggests that the intuition reaction related to fairness norm violations has an innate basis. Our study provides the first evidence that the neural basis underlying this fairness intuition is under genetic control.

In this voxel-wise search for genetic contributions to brain activity in the current study, we find no strong evidence for heritability of regions related to reflective and deliberate process, i.e. integrating both self-interest and a fairness norm to regulate the intuitive system to permit more flexible decision-making (Sanfey *et al.*, 2006; Feng *et al.*, 2015). This finding contradicts previous speculation on the heritability of System 2 based on its close relationship with genetically determined general intelligence and working memory (Evans, 2008), which is under genetic control (Devlin *et al.*, 1997; Blokland *et al.*, 2011). However, the impact of general intelligence and working memory on UG performance has been sparsely addressed in the literature. General intelligence was not predictive of reciprocity outcomes when investigating whether individuals would reciprocate 'generosity' shown from other proposers (Ben-Ner *et al.*, 2004), and cognitive abilities including working memory were not significant in predicting the responder's rejection behavior in the UG (Nguyen *et al.*, 2011). The lack of contribution of general intelligence and working memory in the UG may account for the current observation that no significant genetic contributions to the activity of regions related to reflective and deliberate process.

Culture and gene co-evolution models (Boyd and Richerson, 1985; Gintis, 2003; Fehr and Fischbacher, 2004) provide a theoretical account for our finding that psychological processes supported by the AI are moderately heritable. This dual-evolution model posits that human evolution of social norms has been substantially influenced by the interaction of our cultural and genetic inheritance systems (Chudek and Henrich, 2011). Computational model studies also suggest that fairness is a product of natural selection and that compliance with a fairness norm has advantages in evolution (Young, 1993; Ellingsen, 1997; Nowak and Sigmund, 2005; Rand *et al.*, 2013). This intuition to fairness norm violation may be hardwired into human nature

by natural selection, while culture (or experience) shapes our social behavior with its interaction with genes, and thus the costly punishment behavior is partly under genetic control.

There are several limitations to this study. Although twin studies can suggest that brain activity induced by fairness is partially hardwired, more efforts are needed to identify specific genes in charge of this brain process. Second, the current study only focused on brain activity induced by fairness, which may not causally determine the responder's choice. Although the seminal work of Sanfey found a correlation between the anterior insula and acceptance rate (Sanfey *et al.*, 2003), no study has provided evidence for a causal role of the AI in costly punishment (Gabay *et al.*, 2014; Gu *et al.*, 2015). Third, the neural components of System 1 and System 2 may interact with each other to yield costly punishment; future studies need to investigate the functional interaction between the neural components of System 1 and System 2 using functional or effective connectivity, such as dynamic causal modeling (Friston *et al.*, 2003).

In summary, this study provides evidence for genetic contributions to costly punishment of the responder and its neural basis during the UG. The genetic factor influences the brain activity evoked by unfair proposals in the bilateral insular cortices, suggesting the detection of fairness norm violation is partially hardwired into our brain. Our findings shed more light on the brain processes underlying costly punishment and provide an additional level of evidence for the discussion of the motives underlying this behavior.

Funding

This research was supported by the National Natural Science Foundation of China (91432302, 81371476, 81771473, 71761167001 and 31671166), the National Basic Research Program of China (973 program) (2011CB711002), the National High Technology Research and Development Program of China (863 program) (2015AA020513), Youth Innovation Promotion Association of Chinese Academy of Sciences (2012075) and Beijing Nova Program (Z121107002512064).

Acknowledgments

The authors thank the twins for participating in this study and Jie Zhang and other staff of the BeTwiSt at the Institute of Psychology, Chinese Academy of Sciences for recruiting these twins. The authors would also like to thank Prof. Karl Friston for his suggestions to use PPMs in the voxel-based genetic modeling analysis and to use the standard summary statistic approach to emulate a HLM when accounting for the dependency between twins in the same pair in fMRI data processing. The authors acknowledge the anonymous reviewers for their insightful comments for improving this manuscript.

Supplementary data

Supplementary data are available at SCAN online.

References

Ben-Ner, A., Putterman, L., Kong, F., Magan, D. (2004). Reciprocity in a two-part dictator game. *Journal of Economic Behavior & Organization*, **53**(3), 333–52.

Blokland, G.A.M., McMahon, K.L., Thompson, P.M., Martin, N.G., de Zubicaray, G.I., Wright, M.J. (2011). Heritability of working memory brain activation. *The Journal of Neuroscience*, **31**(30), 10882–90.

Blount, S. (1995). When social outcomes aren't fair: the effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, **63**(2), 131–44.

Bollen, K.A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, **17**(3), 303–16.

Boyd, R., Richerson, P.J. (1985). *Culture and the Evolutionary Process*, Chicago: University of Chicago Press.

Buckholtz, J.W., Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, **15**(5), 655.

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

Chang, L.J., Sanfey, A.G. (2011). Great expectations: neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, **8**(3), 277–84.

Chang, L.J., Sanfey, A.G., Yarkoni, T., Khaw, M.W. (2013). Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cerebral Cortex*, **23**(3), 739–49.

Chen, J., Li, X., Chen, Z., *et al.* (2010). Optimization of zygosity determination by questionnaire and DNA genotyping in Chinese adolescent twins. *Twin Research and Human Genetics*, **13**(2), 194–200.

Chudek, M., Henrich, J. (2011). Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, **15**(5), 218–26.

Civai, C. (2013). Rejecting unfairness: emotion-driven reaction or cognitive heuristic? *Frontiers in Human Neuroscience*, **7**, 126.

Civai, C., Crescentini, C., Rustichini, A., Rumiati, R.I. (2012). Equality versus self-interest in the brain: differential roles of anterior insula and medial prefrontal cortex. *NeuroImage*, **62**(1), 102–12.

Corradi-Dell'Acqua, C., Hofstetter, C., Vuilleumier, P. (2014). Cognitive and affective theory of mind share the same local patterns of activity in posterior temporal but not medial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, **9**(8), 1175–84.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R.I., Fink, G.R. (2012). Disentangling self-and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. *Social Cognitive and Affective Neuroscience*, **8**(4), 424–31.

d'Acremont, M., Lu, Z.L., Li, X., Van der Linden, M., Bechara, A. (2009). Neural correlates of risk prediction error during reinforcement learning in humans. *NeuroImage*, **47**(4), 1929–39.

Devlin, B., Daniels, M., Roeder, K. (1997). The heritability of IQ. *Nature*, **388**, 468.

Ellingsen, T. (1997). The evolution of bargaining behavior. *The Quarterly Journal of Economics*, **112**(2), 581–602.

Evans, J.S. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, **7**(10), 454–9.

Evans, J.S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, **59**(1), 255–78.

Fehr, E., Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, **8**(4), 185–90.

Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, **114**(3), 817–68.

Feng, C., Luo, Y.J., Krueger, F. (2015). Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis. *Human Brain Mapping*, **36**(2), 591–602.

Friston, K.J., Penny, W. (2003). Posterior probability maps and SPMs. *NeuroImage*, **19**(3), 1240–9.

Friston, K.J., Harrison, L., Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, **19**(4), 1273–302.

Gabay, A.S., Radua, J., Kempton, M.J., Mehta, M.A. (2014). The ultimatum game and the brain: a meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, **47**, 549–58.

Gintis, H. (2003). The hitchhiker's guide to altruism: gene-culture coevolution, and the internalization of norms. *Journal of Theoretical Biology*, **220**(4), 407–18.

Grecucci, A., Giorgetta, C., Van't Wout, M., Bonini, N., Sanfey, A.G. (2012). Reappraising the ultimatum: an fMRI study of emotion regulation and decision making. *Cerebral Cortex*, **23**(2), 399–410.

Gu, X., Wang, X., Hula, A., *et al.* (2015). Necessary, yet dissociable contributions of the insular and ventromedial prefrontal cortices to norm adaptation: computational and lesion evidence in humans. *The Journal of Neuroscience*, **35**(2), 467–73.

Güth, W., Schmittberger, R., Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, **3**(4), 367–88.

Harlé, K.M., Sanfey, A.G. (2007). Incidental sadness biases social economic decisions in the ultimatum game. *Emotion*, **7**(4), 876.

Harlé, K.M., Chang, L.J., van't Wout, M., Sanfey, A.G. (2012). The neural mechanisms of affect infusion in social economic decision-making: a mediating role of the anterior insula. *NeuroImage*, **61**(1), 32–40.

Henrich, J., Boyd, R., Bowles, S., *et al.* (2005). "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, **28**(6), 795–815.

Keuler, M.M., Schmidt, N.L., Van Hulle, C.A., Lemery-Chalfant, K., Goldsmith, H.H. (2011). Sensory over-responsivity: prenatal risk factors and temperamental contributions. *Journal of Developmental and Behavioral Pediatrics*, **32**(7), 533.

Kline, R.B. (1998). Software review: software programs for structural equation modeling: Amos, EQS, and LISREL. *Journal of Psychoeducational Assessment*, **16**(4), 343–64.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, **314**(5800), 829–32.

Li, J., Wang, W., Yu, J., Zhu, L. (2016). Young children's development of fairness preference. *Frontiers in Psychology*, **7**, 1274.

Lieberman, M.D. (2007). Social cognitive neuroscience: a review of core processes. *Annual Review of Psychology*, **58**, 259–89.

Lydecker, J.A., Pisetsky, E.M., Mitchell, K.S., *et al.* (2012). Association between co-twin sex and eating disorders in opposite sex twin pairs: evaluations in North American, Norwegian, and Swedish samples. *Journal of Psychosomatic Research*, **72**(1), 73–7.

Lynch, S.K., Turkheimer, E., D'onofrio, B.M., *et al.* (2006). A genetically informed study of the association between harsh punishment and offspring behavioral problems. *Journal of Family Psychology*, **20**(2), 190.

Martin, N.G., Eaves, L.J., Kearsey, M.J., Davies, P. (1978). The power of the classical twin study. *Heredity*, **40**, 97.

Mazaika, P.K., Hoeft, F., Glover, G.H., Reiss, A.L. (2009). Methods and software for fMRI analysis of clinical subjects. *NeuroImage*, **47**, 58.

Neale, M., Cardon, L. (2013). *Methodology for Genetic Studies of Twins and Families*, Dordrecht, The Netherlands: Kluwer Academic Publishers.

Nguyen, C.M., Koenigs, M., Yamada, T.H., *et al.* (2011). Trustworthiness and negative affect predict economic decision-making. *Journal of Cognitive Psychology (Hove, England)*, **23**(6), 748–59.

Nowak, M.A., Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, **437**, 1291.

Plomin, R., DeFries, J.C., Knopik, V.S., Neiderhiser, J. (2013). *Behavioral Genetics*, 6th edn, New York: Worth Publishers.

Preuschoff, K., Quartz, S.R., Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, **28**(11), 2745–52.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 1281–302.

Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–63.

Rand, D.G., Tarnita, C.E., Ohtsuki, H., Nowak, M.A. (2013). Evolution of fairness in the one-shot anonymous ultimatum game. *Proceedings of the National Academy of Sciences*, **110**(7), 2581–6.

Rao, L.-L., Zhou, Y., Zheng, D., Yang, L.-Q., Li, S. (2018). Genetic contribution to variation in risk taking: a functional MRI twin study of the balloon analogue risk task. *Psychological Science*, **29**(10), 1679–91.

Rilling, J.K., Sanfey, A.G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, **62**, 23–48.

Sanderson, D.J., Good, M.A., Skelton, K., *et al.* (2009). Enhanced long-term and impaired short-term spatial memory in GluA1 AMPA receptor subunit knockout mice: evidence for a dual-process memory model. *Learning & Memory*, **16**(6), 379–86.

Sanfey, A.G. (2007). Social decision-making: insights from game theory and neuroscience. *Science*, **318**(5850), 598–602.

Sanfey, A.G., Chang, L.J. (2008). Multiple systems in decision making. *Annals of the New York Academy of Sciences*, **1128**(1), 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**(5626), 1755–8.

Sanfey, A.G., Loewenstein, G., McClure, S.M., Cohen, J.D. (2006). Neuroeconomics: cross-currents in research on decision-making. *Trends in Cognitive Sciences*, **10**(3), 108–16.

Satpute, A.B., Lieberman, M.D. (2006). Integrating automatic and controlled processes into neurocognitive models of social cognition. *Brain Research*, **1079**(1), 86–97.

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**(4), 339–47.

Wallace, B., Cesarini, D., Lichtenstein, P., Johannesson, M. (2007). Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(40), 15631–4.

Warneken, F., Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, **311**(5765), 1301–3.

Xiang, T., Lohrenz, T., Montague, P.R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, **33**(3), 1099–108.

Young, H.P. (1993). An evolutionary model of bargaining. *Journal of Economic Theory*, **59**(1), 145–68.

Zhou, Y., Wang, Y., Rao, L.L., Yang, L.Q., Li, S. (2014). Money talks: neural substrate of modulation of fairness by monetary incentives. *Frontiers in Behavioral Neuroscience*, **8**, 150.