Neural basis of increased costly norm enforcement under adversity

Yan Wu,[1,2] Hongbo Yu,[2] Bo Shen,[2] Rongjun Yu,[3] Zhiheng Zhou,[2] Guoping Zhang,[2,4] Yushi Jiang,[5] and Xiaolin Zhou[2,6,7]

[1]Department of Psychology, School of Educational Sciences, Hangzhou Normal University, Hangzhou 310036, China, [2]Center for Brain and Cognitive Sciences and Department of Psychology, Peking University, Beijing 100871, China, [3]School of Psychology and Center for Studies of Psychological Application, South China Normal University, Guangzhou 510631, China, [4]The China Academy of Corporate Governance and Business School, Nankai University, Tianjin 300071, China, [5]School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China, [6]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China, and [7]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

Humans are willing to punish norm violations even at a substantial personal cost. Using fMRI and a variant of the ultimatum game and functional magnetic resonance imaging, we investigated how the brain differentially responds to fairness in loss and gain domains. Participants (responders) received offers from anonymous partners indicating a division of an amount of monetary gain or loss. If they accept, both get their shares according to the division; if they reject, both get nothing or lose the entire stake. We used a computational model to derive perceived fairness of offers and participant-specific inequity aversion. Behaviorally, participants were more likely to reject unfair offers in the loss (*vs* gain) domain. Neurally, the positive correlation between fairness and activation in ventral striatum was reduced, whereas the negative correlations between fairness and activations in dorsolateral prefrontal cortex were enhanced in the loss domain. Moreover, rejection-related dorsal striatum activation was higher in the loss domain. Furthermore, the gain–loss domain modulates costly punishment only when unfair behavior was directed toward the participants and not when it was directed toward others. These findings provide neural and computational accounts of increased costly norm enforcement under adversity and advanced our understanding of the context-dependent nature of fairness preference.

Keywords: fairness; costly norm enforcement; ultimatum game; fMRI; computational modeling

INTRODUCTION

The concept of fairness and equity in social interaction has occupied the mind of philosophers (Rawls, 1971; Rousseau, 1754/2011; Aristotle, 1998) and economists (Kahneman et al., 1986; Buchanan, 1987) throughout recorded history. As the political philosopher John Rawls noted, 'the fundamental idea in the concept of justice is fairness' (Rawls, 1958). The maintenance of fairness typically requires some forms of punishment, as certain individuals are inevitably tempted by their self-interest to violate the norm of fairness in social encounters (Sober and Wilson, 1999; Fehr and Gächter, 2002; Montague and Lohrenz, 2007). Indeed, humans are willing to enforce the fairness norm by punishing violations even at substantial personal costs (Boyd et al., 2003; Henrich et al., 2006).

A classical way to elicit inequality aversion and costly norm enforcement in laboratory settings is the 'ultimatum game' (UG) in which one player (the proposer) is endowed with an amount of money (e.g. $10) and proposes a division (e.g. keep $7/offer $3) to another player (the responder). The responder can either accept or reject the offer, with the corresponding consequence of both parties receiving the proposed shares or getting nothing (Güth et al., 1982). Behaviorally, responders routinely reject unfair or unequal divisions of resources, with the rejection rate increasing as the unfairness or the inequality of the offer increases (Camerer, 2003). This behavioral preference for fairness is observed in domestic dogs (Range et al., 2009) and non-human primates (Proctor et al., 2013), suggesting a long evolutionary history to the human sense of fairness. In humans, neuroimaging studies with UG show that unfair division of resources elicits negative affect and interoceptive responses represented in the insula (Sanfey et al., 2003; Güroğlu et al., 2010; Hollmann et al., 2011; Kirk et al., 2011), and that rejection of an unfair offer, and thus punishment of norm violations, activates reward-related brain regions such as the dorsal striatum (DS; de Quervain et al., 2004; Baumgartner et al., 2008). Moreover, rejection of unfair offer usually comes at a personal cost to the punisher and thus requires the inhibition of selfish impulses, subserved by dorsolateral prefrontal cortex (DLPFC; Knoch et al., 2006, 2008; Buckholtz et al., 2008; Baumgartner et al., 2011, 2012).

However, most of the previous studies were conducted in a gain domain, leaving aside the situations in which individuals have to share a certain amount of loss. The latter situations are common in human society, such as the liquidation of a bankrupt company. Are people more or less willing to bear the cost to enforce the fairness norm by punishing violators in liability sharing than in gain sharing? As potential losses tend to have a greater impact than equivalent gains upon individuals' choices (Tversky and Kahneman, 1981, 1991; Tom et al., 2007), fairness preference and costly norm enforcement in liability sharing might not be the same as in gain sharing. In a recent study, we (see also Buchan et al., 2005; Leliveld et al., 2009; Zhou and Wu, 2011) extended the classic UG to the loss domain, in which participants had to decide whether to accept or to reject a division of a certain amount of monetary loss [e.g. 10 monetary units (MUs)] proposed by an anonymous partner in a one-shot manner. As in the classic UGan cla otherwise both sides had to pay the whole price (i.e. losing 10 U each). We found that participants were more likely to reject unfair offers in the loss than in the gain domain, suggesting a higher propensity of norm enforcement under adversity. This finding cannot be explained solely by strategic comparisons in making decisions between the two domains as this effect was present regardless of whether the gain–loss frame was manipulated within- or between-participants (Zhou and Wu, 2011).

Two possible motives may underlie this behavioral pattern. One is 'a legitimate passion for equality that incites men to want to be strong and esteemed' and the other 'a depraved taste for equality . . . that leads the weak to want to bring the strong down to their level' (Tocqueville, 1835/2010). According to the first theory, costly norm enforcement is driven by fairness preference, i.e. 'something equal should be given to those who are equal' (Aristotle, 1998; Fehr and Schmidt, 1999). Neurally, it has been demonstrated that the preference for fairness is implemented in the brain valuation system (Bossaerts *et al.*, 2009; Bartra *et al.*, 2013), most notably the ventral striatum (VS) (Tabibnia *et al.*, 2008; Tricomi *et al.*, 2010). The increased demand for fairness and the increased costly norm enforcement under adversity may therefore be associated with enhanced subjective value of fairness and its representation in VS.

Alternatively, costly norm enforcement may be driven by satisfaction through the punishment of norm violations (Dufwenberg and Kirchsteiger, 2004). Punishing those who have behaved unfairly is intrinsically rewarding to the punisher (Seymour *et al.*, 2007; Espin *et al.*, 2012). This motive is especially strong when the potential punisher is the victim of the norm violations. In this case, punishment is likely to be driven by the desire for revenge or retaliation. Neuroimaging studies have observed activation of DS when punishing breaches of trust (de Quervain *et al.*, 2004), unfair divisions of resources (Sanfey *et al.*, 2003; Baumgartner *et al.*, 2011; Strobel *et al.*, 2011; Crockett *et al.*, 2013) and aggressive behavior (Krämer *et al.*, 2007). Thus, the loss frame may increase costly norm enforcement by enhancing the satisfaction of punishment and associated neural processes in DS.

These two theories provide different predictions about punishment behaviors in different social contexts. Retaliation motivates punishment of unfair behavior inflicted on oneself (second-party punishment) but not on others (third-party punishment), whereas fairness preference motivates punishment of unfair behavior directed both toward oneself and toward others (Fehr and Fischbacher, 2004). Therefore, if gain–loss context regulates fairness preferences, loss should increase both second- and third-party punishment, and neuroimaging should reveal enhanced fairness-related responses in the medial prefrontal cortex (MPFC) and VS, which are related to the social utility of fairness (Tabibnia *et al.*, 2008; Tricomi *et al.*, 2010). On the other hand, if gain–loss context regulates retaliatory motives, loss should increase second-, but not third-party punishment and enhance the responses in the DS, anterior cingulate cortex and insula, which are related to reactive aggression and retaliation (Sanfey *et al.*, 2003; de Quervain *et al.*, 2004; Crockett *et al.*, 2013; White *et al.*, 2014).

Here, combining our previous behavioral paradigm with functional magnetic resonance imaging (fMRI), we investigated the neural basis of increased costly norm enforcement by recording the brain hemodynamic responses while participants, as responders in the UG game, were exposed to fair/unfair offers in both loss and gain domains (see also Guo *et al.*, 2013). Using a formal inequality aversion model (Messick and McClintock, 1968; Fehr and Schmidt, 1999), we obtained subjective utility (SU) associated with each offer for each individual participant and a parameter (i.e. the 'envy' parameter) that quantifies how much a particular participant cares about inequality. We then applied these parameters to fMRI data (Wright *et al.*, 2011). A key advantage of using model-based approach over more conventional neuroimaging approaches is that the former can 'provide insights into how a particular cognitive process is implemented in a specific brain area as opposed to merely identifying where a particular process is located' (O'Doherty *et al.*, 2007). Moreover, the model parameters have more specific psychological meanings as compared with rejection rates (Messick and McClintock, 1968; Fehr and Schmidt, 1999). In our imaging analysis, brain activity was time-locked to the presentation of the offers. Based on previous studies, we hypothesized that the mesolimbic areas, such as VS and ventromedial prefrontal cortex (VMPFC), would be activated by monetary gain and fairness (Tabibnia *et al.*, 2008; Haber and Knutson, 2010; Tricomi *et al.*, 2010). DLPFC, anterior cingulate cortex (ACC) and insula, in contrast, would be activated by monetary loss and inequality (Sanfey *et al.*, 2003; Baumgartner *et al.*, 2011, 2012; Chang and Sanfey, 2013). Critically, we predicted that the gain–loss frame would modulate the neural processing of monetary interest and fairness implemented in these brain structures, which may give rise to the behavioral observation of increased costly norm enforcement under adversity.

We also conducted a behavioral experiment to differentiate to what extent the brain activations observed in the fMRI experiment were related to the general fairness preference or to the desire of retaliation. In particular, we asked participants to act as a third-party who could use their own endowment to punish the unfair proposers by reducing the proposers' payoff at a 1:5 ratio. Given that it is established that the third-party punishment is primarily driven by fairness preferences (Fehr and Fischbacher, 2004), this experiment allowed us to address the crucial question as to whether the increased costly punishment in the loss domain is due to fairness preference or desire of retaliation.

METHODS

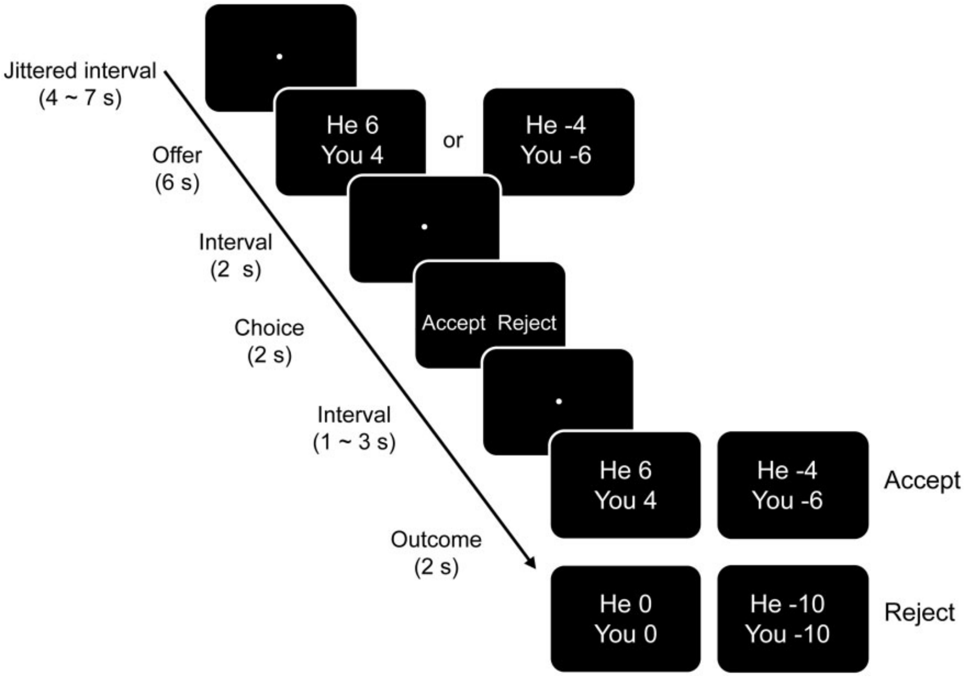
Participants

Twenty-eight healthy volunteers participated in the neuroimaging study in return for payment. Ten participants were excluded from data analysis either due to excessive head motion (>3 mm; three persons) or pure rational response (accepting essentially all offers; seven persons). The remaining 18 (13 females) participants had a mean age of 21.6 ± 1.8 years. Another 31 volunteers (18 females; mean age: 21.0 ± 1.4 years) participated in the behavioral third-party punishment experiment. All the participants were right-handed and were screened for psychiatric or neurological problems. The experiment protocols were approved by the Ethics Committee of the Department of Psychology at Peking University. All participants provided written informed consent after the procedures had been fully explained, in accordance with the Declaration of Helsinki.

Design and procedures

Prior to the experiment, participants were informed of the rules for the UG game. They were told that they would play the game with a group of anonymous students whom they did not know and who had made proposals, in a pilot session, on how to share a 10 MU gain or loss between them and the participants. In each round, the participants and a randomly drawn proposer would be endowed with 10 MUs (in the gain domain) or had to pay for a 10 MU penalty (in the loss domain); the participant's task was to either accept or reject the offer made by the proposer. Upon acceptance, the gain or loss was split as proposed. Upon rejection, the participants and the proposer would receive nothing for the round of endowment or would pay out 10 MUs each for the round of penalty. The participants were informed that they would get 60 Chinese yuan for participating in the fMRI scanning and an additional 0–40 yuan for bonus, the amount of which was dependent on their performance in the experiment. They were also led to believe that the games would be played for real with the set of proposers they

Downloaded from http://scan.oxfordjournals.org/ at Peking University on February 7, 2014



. 1 Sequence of events and timing in a trial. Each trial began by presenting the offer to the participant for 6 s. The participant was told to evaluate the offer but not to press any button at this moment. After a 2 s interval, the participant had to decide whether to accept or to reject the offer by pressing one of two buttons. After another interval, the duration of which varied from 1 to 3 s, the outcome of this trial was presented. Upon acceptance, the amount of gain or loss would be divided according to the proposer's offer. Upon rejection, both the participant and the proposer would get nothing (in the gain domain) or have to pain the full price (in the loss domain).

encountered during the experiment and that their decision in each trial was directly related to their own and the corresponding proposer's final payoff. Participants were debriefed and thanked before they left the testing room.

Unknown to the participants, the offer in each round was predetermined by the experimenter. A two (domain: gain loss) by five (fairness level: 5:5, 4:6, 3:7, 2:8, 1:9) factorial design was used. Division schemes were 5/5, 4/6, 3/7, 2/8, 1/9 for the gain domain, and −5/−5, −6/−4, −7/−3, −8/−2 and −9/−1 for the loss domain, with the number before the slash indicating the offered amount to the responder and the number after the slash indicating the amount left to the proposer. Prior to scanning, each participant was familiarized with the task through 20 out-of-scanner practice trials which had the same composition of experimental conditions as the formal experiment in the scanner.

At the beginning of each trial, a fixation dot was presented at the center of the screen for a jittered duration (4–7 s; Figure 1). The participant then saw the offer (e.g. 'you 2, he 8' in the gain domain, and 'you −8, he −2' in the loss condition). The offer screen remained on the screen for 6 s, during which the participants evaluated the offer without making a response. Then the offer screen disappeared and the fixation dot reappeared for another 2 s, followed by a response screen also presented for 2 s. This screen, with the word 'Accept' on the left and the word 'Reject' on the right, counterbalanced over participants, prompted the participant to make either an acceptance or rejection decision by pressing a corresponding button using the right index or middle finger of the right hand. After a varied duration of 1–3 s, the outcome for the participant and the proposer (e.g. 'you −10, he −10') was displayed for 2 s (Figure 1). Thus, on average, a single round lasted for 18 s. There were 14 trials for each type of offer, with the 10 types of offers being presented in pseudorandom sequence for each participant. The 140 trials were divided into two equal-length runs, resulting in a total scanning time of ~45 min. Each type of offer was equally divided into the two sessions.

In the third-party punishment task, participants observed multiple rounds of one-shot UG. They were told that a number of players had participated in an UG. The responders could either reject or accept the offers by the proposer, or they could give up their right to make a decision. The division schemes for the participants who gave up their decision were presented to the participants who were asked to make third-party decision as to whether to punish the proposers. On each trial they were endowed with 1 yuan, which they could use to reduce the proposer's payoff at a 1:5 ratio. Participants could use 0.2, 0.4, 0.6, 0.8 or 1.0 yuan from their endowment and reduce the proposer's payoff up to 5.0 yuan.

The proposer's division schemes were identical to the second-party punishment paradigm, i.e. 5/5, 4/6, 3/7, 2/8, 1/9 for the gain domain, and −5/−5, −6/−4, −7/−3, −8/−2 and −9/−1 for the loss domain. Each type of division was presented six times in a pseudo-random sequence, rendering 60 critical trials. The participants were instructed that five rounds would be randomly selected and the payoffs to them, the proposer and the responder were computed according to the choices in these rounds. The dependent variable was then the amount of money spent on reducing the proposer's payoff. Since a credible scenario is critical, as it ensures that the third-party experiment and the second-part one can actually be compared, we took three measures to ensure credibility: (i) we excluded any participants who had previously encountered economic games from participation in the study, (ii) we asked each participant after the experiment whether he/she believed the game was real (no participants were suspicious of the scenario) and (iii) we asked each participant to withhold the content of the experiment until we finished data collection (after which, we notified the participant that the data collection was complete and the experiment could be discussed).

fMRI data acquisition

Functional images were acquired on a 3T Siemens Trio system at the Key Laboratory of Cognitive Neuroscience and Learning, Beijing

Downloaded from http://scan.oxfordjournals.org/ at Peking University on February 7, 2014

Normal University, using a T2-weighted echo planar imaging sequence (48 sagittal slices, 3 mm thickness; TR = 2400 ms; TE = 25 ms; flip angle = 90°; field of view = 224 × 224 mm^2 ; voxel size = 3 × 3.5 × 3.5 mm^3). The first five volumes were discarded to account for magnetic equilibration. Two runs of 535 volumes were collected from each participant.

Behavioral modeling

We examined the correspondence between the prediction of a formal economic model and participants' choices (Messick and McClintock, 1968; Fehr and Schmidt, 1999). The aim of this analysis was 2-fold: the first was to explicitly distinguish between the SU of an offer to a particular participant and the degree to which he/she cared about the inequality in that offer. These two psychologically different factors are otherwise implicitly embedded in the participant's choice. The second purpose was to derive model parameters that could bridge the external choices on the one hand and the underlying neural mechanisms on the other (see below).

Following the procedure of Wright al. (2011), we fitted the behavioral data (i.e. the acceptance rate at each fairness level) using a psychometric model,

$$(\text{accept}) = \frac{1}{1 + {}^{-({}_{0} + {}_{1})}},$$

where is the fairness level and $_0$ and $_1$ are free model parameters. We estimated the model separately for the gain and loss domains. Assuming that the acceptance rate is a sigmoid function of SU associated with each offer (Wright al., 2011), the above psychometric equation can be re-written as:

$$(\text{accept}) = \frac{1}{1 + {}^{-\frac{7}{\lambda}}},$$

where the SU (7) is defined according to an influential economic theory of fairness and inequality aversion (Fehr and Schmidt, 1999):

$$7 = {}_{ƀ} - \alpha * (\quad - {}_{ƀ}), \alpha \geq 0; \lambda \geq 0.$$

Instead of denoting direct payoff derived from the offers (e.g. 4 or −8), $_{\text{self}}$ and $_{\text{other}}$ here denote the generalized payoff, i.e. the additional amount of money the proposer ($_{\text{other}}$) and the responder ($_{\text{self}}$) would get when the responder accepts relative to rejects offer. In the gain domain, these values are equal to the proposed division. In the loss domain, these values equal to 10 plus the proposed division (i.e. 1, 2, 3, 4, 5 for the responder and 9, 8, 7, 6, 5 for the proposer). This transformation, while keeping the shape of the regression curves, and thus keeping the model parameters unchanged, aligns the curve in the loss domain with that in the gain domain in a Cartesian two-dimensional space (i.e. generalized payoff as -axis and acceptance rate as -axis) so that a direct comparison between gain and loss is made easy. The 'envy' parameter α reflects the degree to which an individual cares about inequality, and the 'temperature' parameter λ reflects decision randomness. We optimized participant-specific α and λ, separately for gain trials and loss trials, according to the acceptance rate in each condition using the glmfit function implemented in Matlab (Table 1).

1 Results of behavioral model fitting

Participant	Gain			Loss		
	α	λ	Log-likelihood	α	λ	Log-likelihood
1	0.12	0.02	−3.60	0.12	0.02	−5.75
2	0.42	0.74	−18.15	0.94	0.77	−12.15
3	0.53	0.74	−16.30	0.71	0.89	−16.96
4	0.74	0.03	−8.38	0.75	0.03	−9.12
5	0.13	0.02	−3.60	0.73	0.03	−3.60
6	0.33	0.02	−5.74	0.54	0.47	−9.39
7	0.12	0.02	−9.56	0.13	0.02	−5.74
8	0.12	0.02	−5.74	0.33	0.02	−5.74
9	0.12	0.02	−8.38	0.13	0.02	−9.56
10	0.17	0.85	−20.71	0.13	0.02	−7.28
11	0.56	0.81	−17.36	1.11	0.03	−0.00
12	0.33	0.03	−9.12	0.47	0.43	−9.38
13	0.33	0.50	−14.04	0.44	0.73	−17.34
14	0.34	0.03	−3.60	0.34	0.03	−3.60
15	0.74	0.03	−5.75	0.76	0.03	−5.74
16	0.21	0.02	−0.00	0.32	0.93	−23.41
17	0.13	1.11	−24.70	0.20	1.14	−26.02
18	0.34	0.03	−7.27	1.07	0.71	−9.39

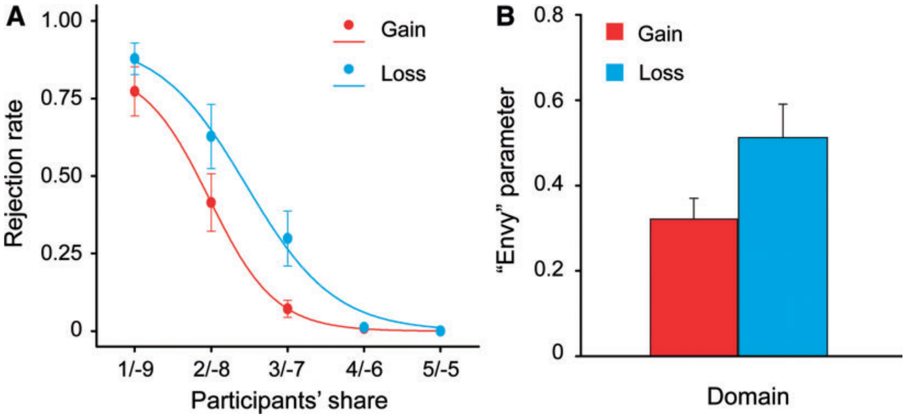
The 'envy' parameter α reflects the degree to which an individual cares about inequality, and the 'temperature' parameter λ

fMRI data analysis

Functional data were analyzed using standard procedures in SPM8 (Statistical Parametric Mapping; http://www.fil.ion.ucl.ac.uk/spm). Images were slice-time corrected, motion corrected, re-sampled to 3 × 3 × 3 mm^3 isotropic voxel, normalized to MNI space (Montreal Neurology Institute), spatially smoothed with an 8 mm FWHM Gaussian filter, and temporally filtered using a high-pass filter with 1/128 Hz cutoff frequency. Statistical analyses based on general linear model (GLM) were performed first at the participant level and then at the group level.

For the individual participant level analysis, we built a parametric model and a factorial model. In the parametric model (GLM 1), we separately modeled the offer presentation, response cue, motor response and outcome in the gain and loss domains with boxcar functions spanning the whole event convolved with a canonical hemodynamic response function. The regressors corresponding to the offer presentation screen in both the gain and the loss domains were further modulated by the estimated SU that was computed with the above modeling procedures. (We also built a parametric model in which the defined fairness level, instead of the SU, served as the parametric modulation. Essentially, the same pattern of activations was obtained.) We checked the correlations between the regressors and found that the correlation between the offer stage and the decision stage was 0.19 and the correlation between the decision stage and the outcome stage was 0.13. These correlations were tolerable high in an event-related fMRI design. In the factorial model (GLM 2), the offer presentation events were assigned to four regressors according to the gain/loss domain and the participants' choice (acceptance rejection). Another six regressors were included corresponding to the onset and duration of the response cue, motor response and outcome in both gain and loss domains. To extract regional activation strength (i.e. beta estimates), a third model was built (GLM 3), in which the offer presentation corresponding to each fairness level was modeled in separate regressors. The six rigid body parameters were also included in all the three models to account for head motion artifact.

For the group level analysis, a full factorial model with the parametric regressors in the gain and loss domains was built. This model allowed us to identify the brain regions that showed differential or similar association with SU in the loss and the gain domains. For the commonalities, we defined a conjunction between the positive effect of SU in the gain and the loss domains, and a conjunction between the negative effect of SU in the gain and the loss domains. For the differential effect, we first defined four contrasts with an exclusive mask approach in parametric analysis (Pochon al., 2002; Seidler al., 2002; Roggeman al., 2011; Chen and Zhou, 2013): (i) positive effect of the parametric regressor of SU in the gain domain exclusively



. **2** Behavioral and modeling results from the UG. (**A**) Points indicate rejection rates in gain (red) and loss domains (blue), respectively. Lines are rejection rates as a logistic function fitted to those points. (**B**) The 'envy' parameter (α) derived from the computational model (Wright *et al.*, 2011) in gain and loss domains. Error bars represent s.e.m.

masked by positive effect of the parametric regressor of SU in the loss domain, i.e. Gain_+ [masked (excl.) by Loss_+], (ii) negative effect of the parametric regressor of SU in the loss domain exclusively masked by negative effect of the parametric regressor of SU in the gain domain Loss_- [masked (excl.) by Gain_-]; (iii) and (iv) the reversed contrast of (i) and (ii), i.e. Loss_+ [masked (excl.) by Gain_+] and Gain_- [masked (excl.) by Loss_-]. The mask image was thresholded at < 0.01 uncorrected. The Gain_+[masked (excl.) by Loss_+] contrast, for example, will show brain areas that positively correlate with SU in the gain domain (at < 0.001) but not positively correlate with SU in the loss domain (even at < 0.01). This difference in significance, however, should not be taken as significant difference (Nieuwenhui *al.*, 2011). For a formal test for significant difference in the association with SU and fairness level, we extracted from two regions of interests (ROIs), i.e. the VS and the right DLPFC, the beta values corresponding to all the 10 offer types (based on GLM 3) and subjected them to repeated measures of analysis of variance (ANOVA). The coordinates of the ROIs were defined based on the exclusive mask procedure. Because the criteria for ROI selection (based on GLM 1) and ROI data extraction and statistical analyses (based on GLM 3) were independent, we believe this procedure controlled for the 'double dipping' problem (Kriegeskorte *al.*, 2009).

To reveal the interaction between gain–loss frame and participants' behavioral choice, the contrast corresponding to this interaction [$\text{Loss}_{\text{(rej–acc)}} - \text{Gain}_{\text{(rej–acc)}}$] was defined using the one sample -test in SPM8 based on GLM 2.

We reported only those clusters that survive cluster-level correction for multiple comparison (family wise error, FWE; < 0.05) either over the whole brain or over *a* ROIs (cluster-level correction after voxel-level thresholding at < 0.005; Lieberman and Cunningham, 2009). The *a* ROI of DS (MNI coordinates: −16, 2, 14) was derived from Crockett *al.* (2013), that of VS (MNI coordinates: −9, 12, −6) and VMPFC (MNI coordinates: −9, 39, −9) were derived from Tricomi *al.* (2010), and that of the anterior cingulate cortex (MNI coordinates: −8, 26, 28) was derived from Sanfey *al.* (2003). Statistical analyses over the ROIs were conducted using the small-volume correction (SVC) method implemented in SPM8. Specifically, spherical search space with 8 mm radius was defined surrounding the independently defined coordinates reported above.

RESULTS

Behavioral results

We first examined the behavioral effects of gain–loss frame on costly norm enforcement in the UG (Figure 2A). Not surprisingly, participants were increasingly likely to reject offers as the level of unfairness increased, Greenhouse–Geisser adjusted $F_{(4,68)} = 61.73$, < 0.001. Importantly, replicating our previous study (Zhou and Wu, 2011), the overall rejection rate was higher for the offers in the loss domain (mean rejection rate = 36.3%, s.d. = 4.1%) than in the gain domain (mean rejection rate = 25.4%, s.d. = 3.4%), $F_{(1,\ 17)} = 35.20$, < 0.001. Moreover, the interaction between domain and fairness level was significant, Greenhouse–Geisser adjusted $F_{(4,\ 68)} = 4.06$, < 0.05, such that the difference in rejection rate tended to be the largest for the moderately unfair conditions (3:7 and 2:8).

Results of behavioral modeling

Although the α parameter was significantly larger than zero in both domains ($\alpha_{\text{gain}} = 0.32 \pm 0.21$, range, 0.12–0.74, $_{17} = 6.60$, < 0.001; $\alpha_{\text{loss}} = 0.51 \pm 0.33$, range, 0.12–1.11, $_{17} = 6.58$, < 0.001), the value was significantly larger in the loss than in the gain domain ($_{17} = 3.37$, < 0.01; Table 1, Figure 2B). The 'temperature' parameter λ ($\lambda_{\text{gain}} = 0.28 \pm 0.09$, range, 0.01–1.11; $\lambda_{\text{loss}} = 0.35 \pm 0.10$, range, 0.02–1.14) did not significantly differ between gain and loss domains ($_{17} < 1$, > 0.1; Table 1). These results confirmed our prediction that people are generally inequality aversive, and more so in the loss than in the gain domain.

Brain areas tracking subjective utility (SU)

Compared with offers in the loss domain, offers in the gain domain elicited activations in the mesolimbic dopaminergic regions (based on GLM 1), including VMPFC, ventral tegmental area (VTA) and posterior cingulate cortex (PCC) (Table 2). The reversed contrast revealed activations in bilateral DLPFC, inferior parietal lobule (IPL) and superior parietal lobule (Table 2). A conjunction analysis showed that in both the loss and the gain domains, VMPFC, VS and PCC were more activated as the utility of the offers increased (Table 3). In contrast, activations in bilateral DLPFC, ACC, anterior insula (AI), thalamus and midbrain negatively correlated with the SU, irrespective of domain (Table 3).

Examining the correlations between brain activations associated with the offer presentation stage and fairness (i.e. SU) of the offers, we found significant correlation in the VS, DS and VMPFC in the gain domain but not in the loss domain (Figure 3A, Table 4). We extracted the beta estimates in VS corresponding to each of the 10 offer type (based on GLM 3) and subjected these values to a formal test for significant interaction. A repeated-measures ANOVA with domain and fairness level as within participant factor showed a significant interaction between domain (gain loss) and fairness level, $F_{(4,\ 68)} = 2.48$, $= 0.05$, confirming the pattern observed in the whole-brain analysis (Figure 3B). It

Downloaded from http://scan.oxfordjournals.org/ at Peking University on February 7, 2014

can be seen from the figure that in the gain domain the VS activation increased with the increase of offer utility, but this trend was not apparent in the loss domain.

2 Brain activations in the gain *vs* loss contrast

Regions	Hemisphere	Max *T*-value	Cluster size (voxels)	Cluster level corrected P_{FWE}	MNI coordinates		
					x	*y*	*z*
Gain–loss							
VMPFC	R	6.85	155	<0.001	6	47	−11
PCC	R	4.37	60	0.019	9	−10	−11
VTA	L	6.73	167	<0.001	−6	−52	16
Loss–gain							
Putamen	R	5.50	52	0.034	30	5	7
DLPFC	L	7.01	138	<0.001	−36	2	34
Rolandic	R	7.01	112	0.001	45	−10	22
IPL/angular	R	5.32	153	<0.001	27	−61	46
SOG	L	6.54	58	0.022	−21	−67	40
Calcarine	L	8.15	554	<0.001	−18	−73	16

SOG, superior occipital gyrus.

3 Brain activations in parametric contrast (conjunction of gain and loss domains)

Regions	Hemisphere	Max *T*-value	Cluster size (voxels)	Cluster level corrected P_{FWE}	MNI coordinates		
					x	*y*	*z*
Increase with SU							
VMPFC	L/R	4.55	181	0.021	6	53	−17
VS	L/R	3.78	10	0.039 [a]	−3	8	−11
Parahippocampus	R	4.02	151	0.044	18	−28	−10
Fusiform	L	5.65	361	<0.001	−33	−28	−19
Precuneus	L	4.09	178	0.022	−9	−55	13
Decrease with SU							
ACC	L/R	3.99	17	0.025 [a]	−6	32	25
DLPFC	R	5.36	246	0.004	39	26	28
	L	5.90	437	<0.001	−42	20	31
Putamen/insula	R	4.58	149	0.047	30	20	1
PAG	R	5.12	278	0.002	3	−22	−14
IPL	R	6.13	1342	<0.001	24	−55	34
	L	5.66			−24	−51	40
IOG	L	4.60	674	<0.001	−42	−76	−8

PAG, periaqueductal gray; IOG, inferior occipital gyrus. [a]SVC based on independently defined ROI (see Methods).

In contrast, we found that the activations in bilateral AI, ACC, right DLPFC, and left lateral orbitofrontal cortex (LOFC) showed negative correlations with fairness (i.e. SU) in the loss domain (Figure 4A, Table 4) but not in the gain domain. The beta estimates (based on GLM 3) in right DLPFC showed a significant interaction between domain (gain loss) and fairness level, $F_{(4,\ 68)} = 2.62$, < 0.05, confirming the pattern observed in the whole-brain analysis (Figure 4B). Specifically, the increase of DLPFC activation in the loss relative to the gain domain was most evident in the most unfair condition (1:9; $_{(17)} = 4.43$, < 0.001), whereas a reversed trend was observed in the fairest condition (5:5; $_{(17)} = -1.82$, $= 0.087$). Moreover, the difference in the mean activation in DLPFC (averaged across all offer levels) in the loss relative to the gain domain predicted the increased rejection rate in the same comparison (robust regression coefficient $= 0.80$, < 0.001; Figure 4C) (Wager al., 2005).

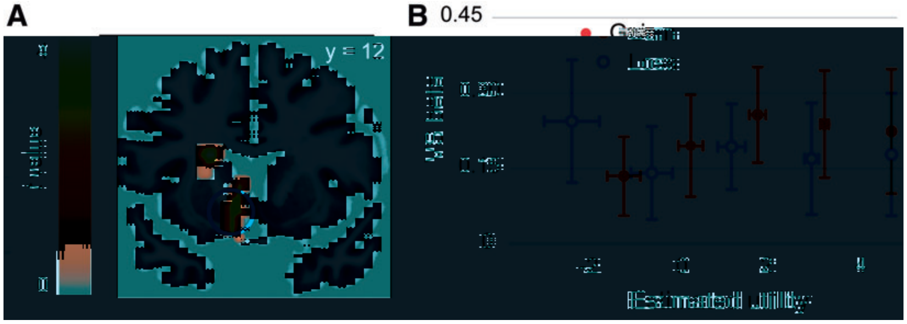
Gain–loss domain modulates rejection-related activation in the dorsal striatum (DS)

To test the hypothesis that loss enhanced the rejection-related activation in the DS, we built a factorial model (GLM 2) which separately modeled the offer presentation events according to the gain/loss domain and the participants' choice (acceptance rejection). Confirming our hypothesis, the contrast 'Loss $_{\text{(rej-acc)}}$ − Gain $_{\text{(rej-acc)}}$' revealed a significant cluster in left DS [MNI coordinates: −15, 2, 22; $_{\text{FWE}}$ (SVC) < 0.05, $= 10$; Figure 5A]. Percent signal change data extracted using Marsbar software (available online at http://marsbar.sourceforge.net/) from a 6 mm radius sphere around the independently defined DS coordinates (Crockett al., 2013) were plotted to illustrate the direction of the interaction (Figure 5B). As can be seen, the interaction was driven by the increased activation during rejection relative to acceptance in the loss domain.

Given that rejection predominantly occurred in the most unfair condition and acceptance in the fairest condition, it may be argued that the observed interaction between participants' choice and domain is driven by the differential effects associated with the two extreme conditions. To test this possibility, we extracted the beta values from the peak voxel of DS and found that the interaction between fairness level (1:9 5:5) and domain was not significant, $F_{(1,\ 17)} < 1$, > 0.1, indicating that the observed effect was not solely due to the differential effect of these two conditions.

We then examined whether increases in DS activation during rejection in the loss domain (gain domain) were correlated with increases in rejection rate. We found that participants with the greatest increases in left DS activation during rejection in the loss domain showed the greatest increases in rejection rate in the loss domain (robust regression coefficient $= 0.67$, < 0.05; Figure 5C).

Downloaded from http://scan.oxfordjournals.org/ at Peking University on February 7, 2014

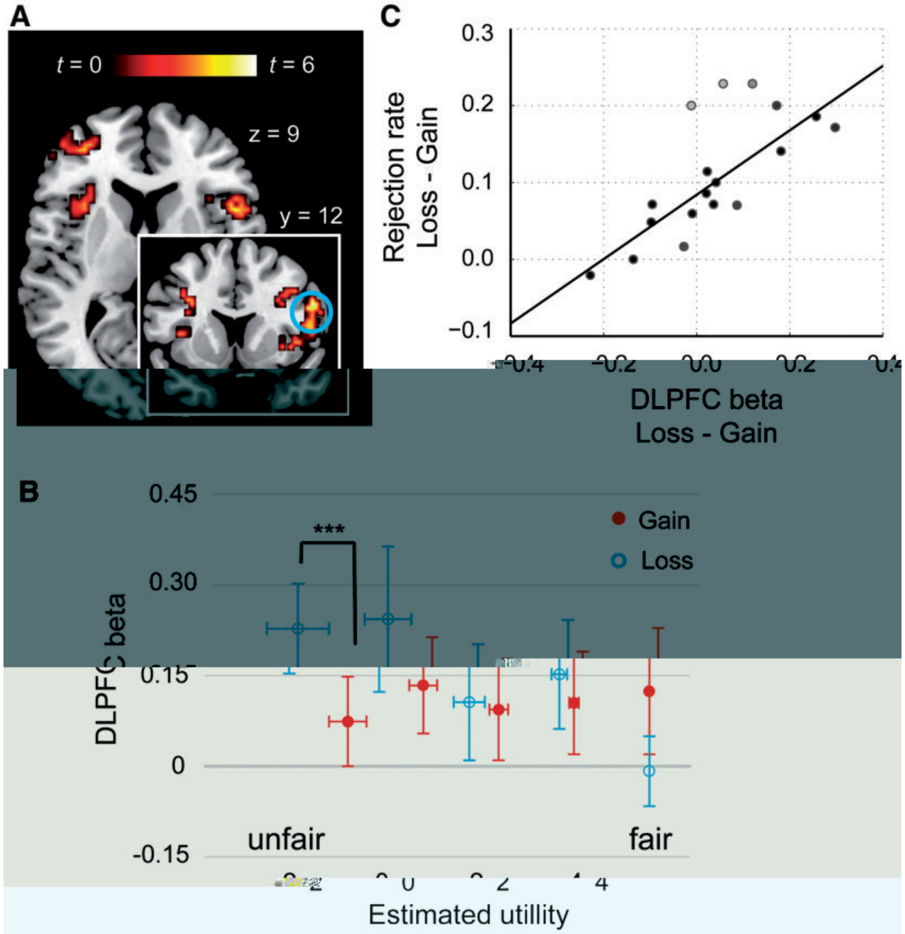


. 3 Positive effect of SU modulated by frame. (**A**) Whole-brain exploratory analysis of the contrast 'Gain$_+$ [masked (excl.) by Loss$_+$]'. (**B**) Beta values corresponding to 10 types of offers (based on GLM 3) extracted from the VS peak. Error bars represent s.e.m.

‐ ‐ **4** Brain activations in parametric contrast (domain-specific activations)

Regions	Hemisphere	Max *T*-value	Cluster size (voxels)	Cluster level corrected P_{FWE}	MNI coordinates		
					x	*y*	*z*
Positive association with SU only in gain domain							
VMPFC	L	5.02	378	<0.001	−6	62	1
Caudate	L	4.62	673	<0.001	−18	14	19
MTG	R	4.23	154	0.041	51	−22	−11
Fusiform	L	5.10	232	0.006	−60	−43	−8
Angular	R	5.16	182	0.020	60	−52	25
Negative association with SU only in loss domain							
LOFC	L	4.93	170	0.027	−48	47	1
DLPFC	R	5.17	428	<0.001	54	11	28
ACC	L/R	3.74	17	0.025 [a]	−6	26	31
Precentral	L	4.35	155	0.040	−33	−1	43
IPL	L	5.46	257	0.003	−48	−25	40

MTG, middle temporal gyrus. [a]SVC based on independently defined ROI (see Methods).

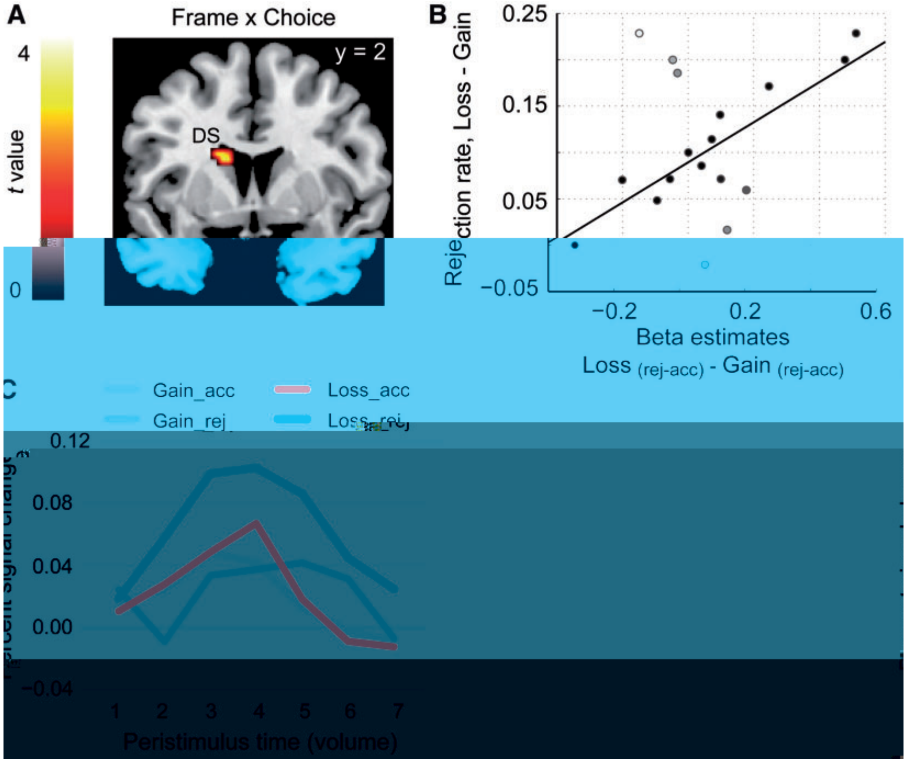


. **4** Negative effect of SU modulated by frame. (**A**) Whole-brain level exploratory analysis of the contrast 'Loss_ [masked (excl.) by Gain_]'. (**B**) Beta values corresponding to 10 types of offers (based on GLM 3) extracted from the DLPFC peak. ( ) The difference in the mean beta values in the gain and loss domain predicted the differences in rejection rates between the loss and gain domain ($r = 0.80$, $P < 0.001$). ***$P < 0.001$ (two-tailed). Error bars represent s.e.m.

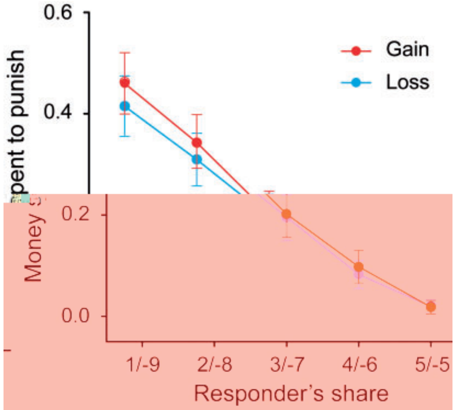
Gain–loss domain and third-party punishment

The neuroimaging findings suggest that the loss domain increased second-party punishment by enhancing retaliatory motives, while at the same time reducing fairness preferences. Lending support to these findings, we found that gain–loss domain did not regulate third-party punishment, which primarily relied on fairness preference rather than retaliatory motives (Figure 6). Specifically, participants paid more to punish proposers as their offers became increasingly unfair (main effect of fairness, $F_{(4,120)} = 48.87$, < 0.001). We did not observe any effects of domain on the amount paid to reduce the proposer's payoff [domain: $F_{(1,30)} = 1.30$, $= 0.264$; domain-by-fairness: $F_{(4,120)} = 1.03$, $= 0.395$].

Downloaded from http://scan.oxfordjournals.org/ at Peking University on February 7, 2014



. 5 Neural effects of interaction between choice and frame. (**A**) ROI-based analysis of the contrast 'Loss (rej-acc) — Gain (rej-acc)'. SVC revealed an activation cluster in the left DS, whose rejection-induced activation was higher in the loss compared with gain domain. (**B**) Activation timecourse extracted from a 6 mm sphere around the maximum coordinates indicates that this interaction effect was driven by the amplified activation difference in the loss relative to the gain domain. ( ) The differences in beta estimates extracted from the activation maximum (Loss — Gain) predicted the increases in rejection rate in the loss relative to the gain domain ($r = 0.67$, $P < 0.05$). Note, the white and grey dots are outliers identified by robust regression and they are down-weighted in computing the correlation coefficients (Wager et al., 2005).



. 6 Effect of fairness and domain on third-party punishment behavior. The amount spent on punishment increased as the offer fairness decreased. Gain–loss domain did not modulate third-party punishment. Error bars represent s.e.m.

DISCUSSION

Using fMRI and a variant of the UG, we provide evidence for a neural and behavioral account of how gain–loss frame modulates costly norm enforcement. Our findings are generally in-line with a recent neuroimaging study (Guo aƀ, 2013) which adopted our previous paradigm (Zhou and Wu, 2011). However, it should be noted that this study focused on the brain responses to aƀ of offers and on the association between brain activations and behavioral measures, such as acceptance rate and subjective value. With the aid of these behavior–brain correlations, we were better able to interpret our neuroimaging results in terms of psychological and economic factors.

Replicating our previous behavioral finding (Zhou and Wu, 2011), participants in the current experiment rejected more in the loss than in the gain domain. Parallel with this, results evidenced a higher response to offers that would be rejected than to those that would be accepted, and critically, the difference was amplified in the loss domain. This raised the possibility that the loss context increased the motivation to reject an unfair offer and thus punish the proposer. Reinforcement learning literature showed that the DS plays a unique role in learning about actions and their reward consequences (Balleine aƀ, 2007). In more complex social context, the DS has been implicated in punishing norm violations (de Quervain aƀ, 2004; Krämer aƀ, 2007; Baumgartner aƀ, 2008, 2012; Strobel aƀ, 2011; Crockett aƀ, 2013). Indeed, Guo aƀ (2013) also found that the rejection (acceptance) of unfair offers elicited higher DS activation in the loss than in the gain domain. Together with these findings, our results suggest that loss may have increased the motivation for punishing norm enforcement, rather than the fairness preferences. This argument is strengthened by the finding that loss did not increase punishment in a third-party context (Figure 6), where punishment was primarily motivated by fairness preferences.

Our results also suggest a context-dependent nature of fairness preference (see also Guo aƀ, 2013). Previous studies have implicated VS in processing the subjective value of fairness and cooperation (Rilling aƀ, 2002; Hsu aƀ, 2008; Tricomi aƀ, 2010). If the loss frame increases costly norm enforcement by enhancing the salience of fairness preferences, we should have observed increased association between VS activation and fairness in the loss relative to the gain domain. But what we observed was just the opposite. Behaviorally, the model output indicated that the same level of unfairness corresponded to lower subjective value in the loss relative to the gain domain.

Downloaded from http://scan.oxfordjournals.org/ at Peking University on February 7, 2014

Neurally, loss has actually blunted the association between the VS activation and fairness. Taken together, our results suggest that loss reduces, rather than enhances the salience of fairness preference.

The increased punishment of norm violations in the loss domain was associated with stronger negative correlations between fairness and the activation in the right DLPFC, which has consistently been implicated in the implementation of second-party punishment such as that in the UG. Existing studies (Knoch *al.*, 2006, 2008) have shown that the DLPFC (especially the right DLPFC) is causally involved in the implementation of costly norm enforcement, perhaps by overriding selfish impulses and thus making the rejection of unfair offers easier (Knoch *al.*, 2006). We found that the negative correlation between right DLPFC activation and fairness increased in the loss relative to the gain domain. The difference between gain and loss in DLPFC activation was most evident in the most unfair conditions, where participants substantially rejected the offers and incurred a cost to

Kahneman, D., Knetsch, J.L., Thaler, R. (1986). Fairness as a constraint on profit seeking: entitlements in the market. *A a E* , *76*, 728–41.

Kirk, U., Downar, J., Montague, P.R. (2011). Interoception drives increased rational decision-making in meditators playing the ultimatum game. *F* , *5*, 49.

Kiser, D., SteemerS, B., Branchi, I., Homberg, J.R. (2012). The reciprocal interaction between serotonin and social behaviour. *a B a a* , *36*, 786–98.

Knoch, D., Nitsche, M.A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., Fehr, E. (2008). Studying the neurobiology of social interaction with transcranial direct current stimulation—the example of punishing unfairness. *C aC* , *18*(9), 1987–90.

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

Krämer, U.M., Jansma, H., Tempelmann, C., Münte, T.F. (2007). Tit-for-tat: the neural basis of reactive aggression. *a* , *38*(1), 203–11.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *a* , *12*, 535–40.

Leliveld, M.C., Beest, I.v., Dijk, E.v., Tenbrunsel, A.E. (2009). Understanding the influence of outcome valence in bargaining: a study on fairness accessibility, norms, and behavior. *J a E a a* , *45*(3), 505–14.

Lieberman, M.D., Cunningham, W.A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *aC a A* , *4*, 423–8.

Messick, D.M., McClintock, C.G. (1968). Motivational bases of choice in experimental games. *J a E a a* , *4*(1), 1–25.

Montague, P.R., Lohrenz, T. (2007). To detect and correct: norm violations and their enforcement. , *56*(1), 14–8.

Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E. (2011). *a* , *14*, 1105–7.

O'Doherty, J., Hampton, A., Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *A a A a* , *1104*, 35–53.

Pochon, J.B., Levy, R., Fossati, P., et al. (2002). The neural system that bridges reward and cognition in humans: an fMRI study. *a aA a 7 a A a*, *99*(8), 5669–74.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *C* , *10*(2), 59–63.

Poldrack, R.A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. , *72*, 692–7.

Proctor, D., Williamson, R.A., de Waal, F.B., Brosnan, S.F. (2013). Chimpanzees play the ultimatum game. *a aA a 7 a A a*, *110*(6), 2070–5.

Range, F., Horn, L., Viranyi, Z., Huber, L. (2009). The absence of reward induces inequity aversion in dogs. *a aA a 7 a A a*, *106*(1), 340–5.

Rawls, J. (1958). Justice as fairness. *a* , *67*(2), 164–94.

Rawls, J. (1971). *A J* . Cambridge, MA: Harvard University Press.

Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., Kilts, C.D. (2002). A neural basis for social cooperation. , *35*(2), 395–405.

Rogers, R.D. (2011). The roles of dopamine and serotonin in decision making: evidence from pharmacological experiments in humans. *a a* , *36*, 114–32.

Roggeman, C., Santens, S., Fias, W., Verguts, T. (2011). Stages of nonsymbolic number processing in occipitoparietal cortex disentangled by fMRI adaptation. *J a* , *31*(19), 7168–73.

Rousseau, J.-J. (1754/2011). *Ba a* . Indianapolis, IN: Hackett Publishing Company.

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. , *300*(5626), 1755–8.

Seidler, R.D., Purushotham, A., Kim, S.-G., Uğurbil, K., Willingham, D., Ashe, J. (2002). Cerebellum activation associated with performance change but not motor learning. , *29*, 2043–6.

Seymour, B., Singer, T., Dolan, R. (2007). The neurobiology of punishment. *a* , *8*(4), 300–11.

Sober, E., Wilson, D.S. (1999). *7 : E a 7 B a* . Cambridge, MA: Harvard University Press.

Strobel, A., Zimmermann, J., Schmitz, A., et al. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *a* , *54*(1), 671–80.

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). *a* , *19*(4), 339–47.

Tocqueville, A. (1835/2010). *D a A a*. Indianapolis, IN: Liberty Fund.

Tom, S.M., Fox, C.R., Trepel, C., Poldrack, R.A. (2007). The neural basis of loss aversion in decision-making under risk. , *315*(5811), 515–8.

Tricomi, E., Rangel, A., Camerer, C.F., O'Doherty, J.P. (2010). Neural evidence for inequality-averse social preferences. *a* , *463*(7284), 1089–91.

Tversky, A., Kahneman, D. (1981). The framing of decisions and the psychology of choice. , *211*(30), 453–8.

Tversky, A., Kahneman, D. (1991). Loss aversion in riskless choice: a reference-dependent model. *a J a E* , *106*(4), 1039–61.

Wager, T.D., Keller, M.C., Lacey, S.C., Jonides, J. (2005). Increased sensitivity in neuroimaging analyses using robust regression. *a* , *26*, 99–113.

White, S.F., Brislin, S.J., Sinclair, S., Blair, J.R. (in press). Punishing unfairness: rewarding or the organization of a reactively aggressive response? *H a B a a* .

Wright, N.D., Symmonds, M., Fleming, S.M., Dolan, R.J. (2011). Neural segregation of objective and contextual aspects of fairness. *J a* , *31*(14), 5244–52.

Zhou, X., Wu, Y. (2011). Sharing losses and sharing gains: increased demand for fairness under adversity. *J a E a a* , *47*(3), 582–8.

Downloaded from http://scan.oxfordjournals.org/ at Peking University on February 7, 2014