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]

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

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 predeter-

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 $3 \times 3.5 \times 3.5$ 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 *a*. (2011), we fitted the behavioral data (i.e. the acceptance rate at each fairness level) using a psychometric model,

$$P \quad \frac{1}{1 \quad b_0 \quad b_1},$$

where is the fairness level and b_0 and b_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 *a*., 2011), the above psychometric equation can be re-written as:

$$P \quad \frac{1}{1 \quad \frac{U}{\lambda}},$$

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

$$U \quad \alpha * \quad ,\alpha \geq 0; \lambda \geq 0.$$

Instead of denoting direct payoff derived from the offers (e.g. 4 or 8), $_{self}$ and $_{other}$ here denote the generalized payoff, i.e. the additional amount of money the proposer ($_{other}$) and the responder ($_{self}$) 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).

Table 1

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 $3 \times 3 \times 3$ 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 *a*., 2002; Seidler *a*., 2002; Roggeman *a*., 2011; Chen and Zhou, 2013): (i) positive effect of the parametric regressor of SU in the gain domain exclusively

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 $P < 0.01$ uncorrected. The Gain [masked (excl.) by Loss] contrast, for example, will show brain areas that positively correlate with SU in the gain domain (at $P < 0.001$) but not positively correlate with SU in the loss domain (even at $P < 0.01$). This difference in significance, however, should not be taken as significant difference (Nieuwenhui *a*., 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 *a*., 2009).

To reveal the interaction between gain loss frame and participants' behavioral choice, the contrast corresponding to this interaction [Loss $_{(\text{rej acc})}$ 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; $P < 0.05$) either over the whole brain or over *a* ROIs (cluster-level correction after voxel-level thresholding at $P < 0.005$; Lieberman and Cunningham, 2009). The *a* ROI of DS (MNI coordinates: 16, 2, 14) was derived from Crockett *a*. (2013), that of VS (MNI coordinates: 9, 12, 6) and VMPFC (MNI coordinates: 9, 39, 9) were derived from Tricomi *a*. (2010), and that of the anterior cingulate cortex (MNI coordinates: 8, 26, 28) was derived from Sanfey *a*. (2003). Statistical analyses over the ROIs were conducted using the small-volume correction (SVC) method implemented in SPM8. Specifically35.5pec6f the anteriorover

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.

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

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 (

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 *a* ., 2013) which adopted our previous paradigm (Zhou and Wu, 2011). However, it should be noted that this study focused on the brain responses to *b c a* 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 (K

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

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

Kiser, D., SteemerS, B., Branchi, I., Homberg, J.R. (2012). The reciprocal interaction between serotonin and social behaviour. *N c c a B b a a R , 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 b a C , 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. *Sc c , 314*(5800), 829 32.

Krämer, U.M., Jansma, H., Tempelmann, C., Münte, T.F. (2007). Tit-for-tat: the neural basis of reactive aggression. *N 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. *Na N c c , 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 S c a P c , 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. *S c a C a A c N c c , 4*, 423 8.

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

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

Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E. (2011). *Na N c c , 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 N Y Aca Sc c , 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. *P c Na a Aca Sc c U S a A ca, 99*(8), 5669 74.

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

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

Proctor, D., Williamson, R.A., de Waal, F.B., Brosnan, S.F. (2013). Chimpanzees play the ultimatum game. *P c Na a Aca Sc c U S a A ca, 110*(6), 2070 5.

Range, F., Horn, L., Viranyi, Z., Huber, L. (2009). The absence of reward induces inequity aversion in dogs. *P c Na a Aca Sc c U S a A ca, 106*(1), 340 5.

Rawls, J. (1958). Justice as fairness. *T P ca R , 67*