Supporting Information for

Neurocomputational evidence that conflicting prosocial motives guide distributive justice

Yue Li[a,b,1] , *Jie Hu*[a,c ,1,2], *Christian C. Ruff* [c], *Xiaolin Zhou*[a,b,d,e,2]

[a]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China

[b]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

[c]Zurich Center for Neuroeconomics, Department of Economics, University of Zurich, Blümlisalpstrasse 10, 8006, Zurich, Switzerland

[d]School of Business and Management, Shanghai International Studies University, Shanghai 200083, China

[e]Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention and School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China

[1] Y.L. and J.H. contributed equally to this work.

[2] To whom correspondence may be addressed:

Dr. Jie Hu, Email: hujie0223@gmail.com, Zurich Center for Neuroeconomics, Department of Economics, University of Zurich, Blümlisalpstrasse 10, 8006, Zurich, Switzerland

Prof. Xiaolin Zhou, Email: xz104@pku.edu.cn, School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China

SI Materials and Methods

Participants

Sixty-three right-handed undergraduate and graduate students were recruited in the experiment. Six participants were excluded because of either making the same decision all the time or excessive head movement (> ± 3 mm in translation and/or > ± 3° in rotation). The remaining 57 participants were aged between 19 and 28 years (mean = 21.83, SD = 1.91; 31 female). Sample size was determined based on previous study (1) that detected a significant rank reversal effect on individuals equality choice with an effect size of d = 0.36 in a laboratory experiment. We thus determined our sample size with G*Power 3.1, which suggested that we need 63 participants to have adequate power ($1 – \beta > 0.80$) to detect an effect with d = 0.36 at the level of $\alpha = 0.05$. No participant reported any history of psychiatric, neurological, or cognitive disorders. Informed written consent was obtained from each participant before the experiment. The study was carried out in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Department of Psychology, Peking University.

Design and procedures

In the present study, we developed a redistribution task to assess individuals' preferences to redistribute unequal wealth allocations. In this task, participants were first presented with a monetary distribution scheme (e.g., Initial offer: Person A: ￥15, Person B: ￥3) between two anonymous strangers. The initial endowment of each party was allocated unequally and randomly by computer, and participants had to choose between two redistribution options (i.e., alternative offers) which transferred a certain amount of money from the one with higher initial endowment (advantaged party) to the one with lower initial endowment (disadvantaged party, Figure 1A). Participants were informed that all of the anonymous strangers in the redistribution task had made the same effort, spent equal amount of time, and made the same contribution in another experiment, and that their decisions would determine those strangers' final payoffs. Moreover, the strangers would only know their own final payoffs but would not know their initial endowments or the payoffs of others.

In the No Rank-reversal condition, both alternative offers (e.g., Offer 1: Person A: ￥14, Person B: ￥4; Offer 2: Person A: ￥10, Person B: ￥8) were more equal than the initial offer (e.g., Person A: ￥15, Person B: ￥3), and both alternative offers kept the same total payoffs and the same relative rankings between the two parties as the initial offer. In the Rank-reversal condition, participants were presented with the same initial offer (e.g., Person A: ￥15, Person B: ￥3) and the same more unequal alternative offer (e.g., Offer 1: Person A: ￥14, Person B: ￥4) as the No Rank-reversal condition, but with a different more equal alternative offer (e.g., Offer 2: Person A: ￥8, Person B: ￥10). This more equal alternative offer (e.g., Person A: ￥8, Person B: ￥10) had the same inequality level as the more equal alternative offer in the No Rank-reversal condition (e.g., Offer 2 in the No Rank-reversal condition: Person A: ￥10, Person B: ￥8), but would reverse the initially relative advantageous/disadvantageous rankings of the two parties (Figure 1B). We matched all trials in the No Rank-reversal condition with the Rank-reversal condition to allow for direct comparison between the two conditions. The difference in the probability of choosing the more equal alternative offer between the No Rank-reversal and Rank-reversal condition was then considered as a behavioral measure of the effect of harm aversion and/or rank reversal aversion on redistribution behavior. To differentiate the effect of inequality and the amount of transferred money (i.e., harm to the advantaged party), we orthogonalized the differences in inequality and the transferred money between the two alternative offers in the Rank-reversal condition (Figure 1D left panel). In addition, we included two filler conditions in which one of the alternative offers was equally distributed (e.g., Person A: ￥9, Person B: ￥9), and the other alternative offer either kept (Filler condition 1, e.g. Person A: ￥14, Person B: ￥4) or reversed (Filler condition 2, e.g. Person A: ￥4, Person B: ￥14) initially relative advantageous/disadvantageous rankings (Figure 1B).

At the beginning of each trial, a fixation point was presented at the center of the screen for 1 s, then the blurred pictures of the two anonymous strangers together with their initial endowments were presented for 3 s. Next, after a blank screen jittering from 1 to 4 s, the two alternative offers were presented. Participants needed to choose one out of the two alternative offers within 6.5 s. After a blank screen jittering from 1 to 4 s, the next trial began (see Figure 1A). The participants knew that 10 trials were randomly selected by computer to determine corresponding persons' final payoffs based on their decisions. There were 66 trials in each of the No Rank-reversal and Rank-reversal conditions, and 15 trials in each filler condition. The 162 trials were divided into three scanning sessions lasting ~15 minutes each. After the experiment, each participant received CNY 120 (~ USD 20) for compensation.

Model-free analysis

We first conducted model-free generalized mixed-effects analysis to test the effects of different components on individuals' probability to choose the more equal alternative offer. In this analysis, we pooled all the trials in the No Rank-reversal and Rank-reversal conditions across all participants. We considered participants' choice as the dependent variable (more equal choice = 1, more unequal choice = 0), and included 1) the absolute value of difference in the initial endowments between the two parties (Δ Initial endowment), 2) the absolute value of difference in the inequality level between the two alternative offers (Δ Inequality), 3) differences in the amount of transferred money between the more equal alternative offer and the more unequal offer (Δ Transfer), 4) condition (Rank-reversal = 1, No Rank-reversal = 0), and interactions between the four variables as the predictors in the model. All these predictors were standardized before being entered into the model and considered as fixed effects, and participants were considered as a random-effect intercept term. We performed this linear mixed-effects analysis using the lme4 package in R.

Computational modeling analysis

To formalize different motives underlying redistribution behaviors, we performed model-based analyses to identify how people weigh between multiple motives to make redistributive decisions. It is noteworthy that we only performed computational modeling analyses for the Rank-reversal condition.

This is because, first, to match with the trials in the Rank-reversal condition, we generated alternative offers in the No Rank-reversal condition with the same inequality levels as the counterpart offers in the Rank-reversal condition and there was no variance in efficiency across alternative offers since the initial offer and both alternative offers had the same sum of payoffs. Therefore, computational models given the trial set in the No Rank-reversal condition could not effectively measure different levels of inequality aversion across participants. Second, as the difference in inequality and the difference in transferred money between alternative offers were completely correlated with each other in the No Rank-reversal condition, it is impossible to differentiate inequality aversion from harm aversion in the No Rank-reversal condition.

We established four families of computational models to formally examine how inequality aversion, harm aversion, and rank reversal aversion affect individuals' redistribution behaviors in the Rank-reversal condition. Different models held different assumptions about how people discounted the utility of the more equal alternative offer [$U(Equal)$] in the Rank-reversal condition.

Model M1 assumed that participants' choices are only influenced by inequality aversion. We followed the classical inequality aversion model proposed by Fehr and Schmidt (1999) in which people assign values to the outcomes of all parties but devalue the inequality they experience for any kinds of distribution. Since both parties of the distribution are anonymous strangers for the participant, we considered the absolute value of the payoff difference between the two parties as the inequality level of the two offers:

$$U(Unequal) = I_A + I_B - \alpha\,|I_A - I_B| \quad (1)$$

$$U(Equal) = E_A + E_B - \alpha\,|E_A - E_B| \quad (2)$$

$$\Delta U = U(Equal) - U(Unequal) = [E_A + E_B - \alpha\,|E_A - E_B|] - [I_A + I_B - \alpha\,|I_A - I_B|]$$
$$= \alpha\,(|I_A - I_B| - |E_A - E_B|) \quad (3)$$

where $I_A(I_B)$ is the payoff of the more unequal alternative offer for initially advantaged (disadvantaged) party, $E_A(E_B)$ is the payoff of the more equal alternative offer for initially advantaged (disadvantaged) party, and α is the inequality aversion parameter that captures the weighing of inequality level of the offers. Since in the current paradigm, the two alternative offers have the same payoff sum ($I_A + I_B = E_A + E_B$), the utility difference (ΔU) is mainly driven by the difference in inequality level between the two offers (i.e., $|I_A - I_B| - |E_A - E_B|$). We refer to this inequality difference as ΔF ($\Delta F = |I_A - I_B| - |E_A - E_B|$). Therefore,

$$\Delta U = \alpha \Delta F \qquad \text{(M1)}$$

Another possibility is that, since only the more equal alternative offer will reverse the relative rankings between the two parties in the initial offer, participants who are averse to rank reversal will devalue the utility of the more equal offer ($U(Equal)$) for rank reversal. Therefore, Model M2 quantified additional effects of rank reversal aversion on top of inequality concerns. We included one discounting parameter δ to capture rank reversal aversion for the more equal offer:

$$U(Equal) = E_A + E_B - \alpha\, |E_A - E_B| - \delta \qquad (4)$$

Then,

$$\begin{aligned} \Delta U = U(Equal) - U(Unequal) &= [E_A + E_B - \alpha\, |E_A - E_B| - \delta] - [I_A + I_B - \alpha\, |I_A - I_B|] \\ &= \alpha\, (|I_A - I_B| - |E_A - E_B|) - \delta \\ &= \alpha \Delta F - \delta \qquad \text{(M2)} \end{aligned}$$

A third possibility is that people may also be averse to benefit one party by harming the other one. Therefore, it is plausible that the more money the alternative offer takes away from the initially advantaged party, the more averse the participant is to the offer. In other words, people will not only devalue the offers for the inequality level and rank reversal but also devalue them for the extent the offers harm the initially advantaged party. To scrutinize how people weigh the harm of alternative offers to the initially advantaged party, we constructed six more models assuming different strategies of devaluing harms.

First, in Model M3a to M3c, we assumed that participants would first evaluate the inequality level of the two alternative offers. Then, we assumed that people would discount the utility of the alternative offer by the amount of money transferred from the initially advantaged party to the disadvantaged party. To reach the more equal offer, participants need to transfer a larger amount of money from the initially advantaged party to the disadvantaged party than to reach to the more unequal offer. Therefore, we assumed that in addition to the difference in inequality level ($\mathit{\Delta F}$) and rank reversal, participants would also consider the difference in the amount of money transferred between the two parties (*ΔT):*

$$U(Unequal) = \mathrm{I}_A + I_B - \alpha\,|\mathrm{I}_A - I_B| - \beta(D_A - I_A) \tag{5}$$

$$U(Equal) = E_A + E_B - \alpha\,|E_A - E_B| - \beta(D_A - E_A) - \delta \tag{6}$$

$$\mathit{\Delta U} = U(Equal) - U(Unequal) = \alpha(\,|I_A - I_B| - |E_A - E_B|) - \beta((D_A - E_A) - (D_A - I_A)) - \delta \tag{7}$$

$$\mathit{\Delta F} = |I_A - I_B| - |E_A - E_B| \tag{8}$$

$$\mathit{\Delta T} = I_A - E_A \tag{9}$$

$$\mathit{\Delta U} = \alpha\mathit{\Delta F} - \beta\mathit{\Delta T} - \delta \tag{M3a}$$

where $D_A(D_B)$ is the payoff of the initial offer for advantaged (disadvantaged) party, $\mathit{\Delta F}$ is the difference in inequality level between the two alternative offers as above models, and $\mathit{\Delta T}$ is the difference in the amount of money transferred across the two parties between the two alternative offers. α and δ are still the inequality aversion parameter and rank reversal aversion parameter as control models (M1 and M2). β is the harm aversion parameter that captures the subjective cost to take money away from the initially advantaged party.

In model M3b, we assumed that people did not consider rank reversal generated by the more equal offer:

$$U(Equal) = E_A + E_B - \alpha\,|E_A - E_B| - \beta(D_A - E_A) \qquad (10)$$

Therefore,

$$\varDelta U = \alpha \varDelta F - \beta \varDelta T \qquad \text{(M3b)}$$

Model M3c assumed that participants discounted the utility of the more equal alternative offer for both transferred money and rank reversal, but that the harm aversion parameter captured additive effects of the transferred money and rank reversal effect:

$$\varDelta U = \alpha \varDelta F - \beta(\varDelta T + \delta) \qquad \text{(M3c)}$$

The models listed above assumed that people would discount the more equal alternative offer for the transferred money. However, it is possible that people are not averse to transfer more money as long as the transferred money can decrease the initial inequality level. Instead, they may be only averse to reach a certain equality level by transferring more money than necessary. Therefore, in models M4a to M4c, we assumed that participants would first evaluate the inequality difference between alternative offers ($\varDelta F$) and rank reversal. Since in the Rank-reversal condition, the more unequal alternative offer decreased the inequality level of the initial distribution and did not reverse the initially relative rankings or transferred more money than necessary, the model assumed that the utility of the more unequal alternative offer was only devalued by the inequality level. For the more equal offer, we assumed that in addition to inequality level, participants would also discount $U(Equal)$ for the proportion of the transferred money that exceeded the necessary amount of money that could reach the same equality level as the equal offer itself but keep the initially relative rankings (i.e., H), which was also considered as the amount of extra harm to the advantaged party. Therefore, model M4a is as follows:

$$U(Unequal) = \mathrm{I}_A + I_B - \alpha\,|\mathrm{I}_A - I_B| \qquad (1)$$

$$U(Equal) = E_A + E_B - \alpha\,|E_A - E_B| - \beta\big((D_A - E_A) - (D_A - E_B)\big) - \delta \qquad (11)$$

$$\Delta U = U(Equal) - U(Unequal)$$

$$= [E_A + E_B - \alpha\,|E_A - E_B| - \beta((D_A - E_A) - (D_A - E_B)) - \delta] - [\mathrm{I}_A + I_B - \alpha\,|\mathrm{I}_A - I_B|]$$

$$= \alpha(\,|I_A - I_B| - |E_A - E_B|) - \beta(E_B - E_A) - \delta \qquad (12)$$

$$\Delta F = |I_A - I_B| - |E_A - E_B| \qquad (8)$$

$$H = E_B - E_A \qquad (13)$$

$$\Delta U = U(Equal) - U(Unequal) = \alpha\Delta F - \beta H - \delta \qquad \text{(M4a)}$$

where α is the inequality aversion parameter that captures the weighing of inequality difference between the two alternative offers, β is the harm aversion parameter that captures the subjective cost of taking more money than necessary from the advantaged party (i.e., generating greater others' loss or harm), δ is the rank reversal aversion parameter, and H is the proportion of the transferred money that exceeds the necessary amount of money that can both reach the same equality level as the equal offer itself and keep the initially relative rankings.

In models M4b and M4c, we also assumed different ways to account for individuals' aversion to rank reversal:

$$\Delta U = \alpha\Delta F - \beta H \qquad \text{(M4b)}$$

and

$$\Delta U = \alpha\Delta F - \beta(H + \delta) \qquad \text{(M4c)}$$

To summarize, in total, we established eight models in four families which held different assumptions about how people devalue the utility of alternative offers in the Rank-reversal condition. In the control models (M1 and M2), we only considered inequality aversion (M1) or additional rank reversal aversion (M2). For the two model families considering harm aversion (M3a – M3c and M4a – M4c), models within the same model family set shared the same way of calculation of harm, but assumed different types of devaluations of harm and rank reversal.

Similar to the notion of the harm signals (i.e., *H*) calculated above, it is also possible that participants would devalue the more equal alternative offer only by the proportion of the transferred money that reversed the initial relative rankings (i.e., $R = (E_B - E_A)/2$), but not by the proportion that reduced the inequality level to the absolute equality level. This psychological component looks differently from the harm signals in M4a to M4c at the first glance but is just double the harm signal as defined above (i.e., $H \ = 2 \cdot R$). Therefore, we did not set up a separate model for this possibility.

For all models, we calculated trial-by-trial utility differences (*ΔU*) between the two alternative offers [$(\Delta U = U(Equal) - U(Unequal)$] and employed a softmax function to transform these utility differences into probabilities of choosing the more equal alternative offer:

$$P(Equal) = \frac{1}{1 + e^{-\lambda\Delta U}}$$

where λ is a free temperature parameter reflecting to what extent an individual's decisions depend on ΔU.

We obtained best fitting parameters by maximizing the log likelihood of the data for each model with the MATLAB function fmincon. To avoid the optimization getting stuck in local minima, we used multiple starting points. To evaluate model fits, we calculated the Bayesian Information Criterion (BIC) (2)which rewards model parsimony to avoid overfitting:

$$BIC = -2 \ln L + k \ln{(n)}$$

where L is the maximized likelihood for the model, k is the number of free parameters in the model, and n is the number of observations. Models were estimated across all participants for model comparison. We also followed established procedures (2) to calculate Bayes factor as $\mathrm{BF} = \exp(-\frac{1}{2}\,\Delta BIC)$, where ΔBIC is the difference in BIC between the winning model (M4a) and each alternative model. BF between 3 and 10 indicates moderate evidence, BF > 10 indicates strong evidence, and BF > 100 indicates very strong evidence that the winning model is superior to the alternative model (2).

Parameter recovery

We performed parameter recovery to validate that the winning model can identify each parameter. We focused on how well the model can recover the three critical parameters: inequality aversion α, harm aversion β, and rank reversal aversion δ. Specifically, we generated 27 datasets using all combinations of three plausible values for each parameter (α: 0.1, 0.3, 0.6; β: 0.1, 0.3, 0.6; δ: 0.8, 1.1, 1.4), with the temperature parameter λ fixed at 1.2. For each parameter combination, we applied the set values of the three parameters to the winning model to simulate agent's responses in the Rank-reversal condition 150 times, and then re-estimated the three parameters for the simulated responses using the winning model to get 150 sets of recovered estimates. We checked how well the distributions of the recovered estimates fit with the true values of the parameters.

Model recovery

We performed model recovery to test how well the data generated by model M3a and M4a can be recovered by both models. Specifically, we generated 54 datasets using all combinations of three plausible values for each parameter (α: 0.1, 0.3, 0.6; β: 0.1, 0.3, 0.6; δ: 0.1, 0.3, 0.6), and the temperature parameter λ as 1.0 and 1.4. For each parameter combination, we applied the set values of the four parameters to the M3a and M4a to generate agent's responses in the Rank-reversal condition, and then re-estimated the parameters for the generated responses and re-generated choices based on the re-estimated parameters for M3a and M4a, respectively. We examined the accuracy of the model simulation by comparing the re-generated choices with the originally generated choices for each model.

Cross-validation prediction analyses and posterior predictive checks

To further evaluate the performance of the winning model, we performed supplementary analyses. First, we did cross-validation prediction analyses by estimating model parameters with each participant's half trials (i.e., odd-numbered trials), and simulating responses with the estimated parameters on the other half trials (i.e., even-numbered trials). Then, we calculated the cross-validated prediction accuracy by comparing simulated responses with observed responses. Second, we simulated each participant's responses by applying their own parameters estimated from all trials to the winning model to generate 100 sets of simulated responses for each participant. Then, we calculated the probability to choose the more equal offer in these 100 sets of responses as simulated behavior and correlated it with observed probabilities of choosing the more equal offer. Note that, to avoid potential effects biased by outliers, we employed nonparametric tests (i.e., Kendall's tau) for all correlation analyses in the current study. Robust regression analyses (with the robustfit function in matlab) were also performed to confirm the correlation relationships after controlling outliers.

Model simulation

We performed model simulation analyses to test whether and how simulated choice would vary with the three parameters of interest in the winning model (i.e., inequality aversion α, harm aversion β, and rank reversal aversion δ). Similar as the parameter recovery analysis, we generated 27 datasets using all combinations of three plausible values for each parameter (α: 0.1, 0.3, 0.6; β: 0.1, 0.3, 0.6; δ: 0.8, 1.1, 1.4). The temperature parameter λ was fixed at 1.2. We simulated agent's responses with the winning model in the Rank-reversal condition 50 times for each parameter combination; and, for each repeat, a small noise derived from a uniform distribution (0, 0.1) was added to each of the three the parameter values. The simulation result is shown in Figure S3.

fMRI data acquisition and preprocessing

We collected T2*-weighted echo-planar images (EPI) using a GE-MR750 3.0T scanner with a standard head coil at Tongji University, China. The images were acquired in 40 axial slices parallel to the AC-PC line in an interleaved order, with an in-plane resolution of 3mm × 3mm, a slice thickness of 4 mm, an inter-slice gap of 4 mm, a repetition time (TR) of 2000 ms, an echo time (TE) of 30 ms, a flip angle of 90°, and a field of view (FOV) of 200mm × 200 mm. We used Statistical Parametric Mapping software SPM12 (Wellcome Trust Department of Cognitive Neurology, London, UK) which was run through Matlab (Mathworks) to preprocess the fMRI images. For each session, the first five volumes were discarded to allow for stabilization of magnetization. For the remaining images, we first performed slice-time correction to the middle slice, then realigned the images to account for head movement, spatially re-sampled the images to 3 × 3 × 3 isotropic voxel, normalized them to standard Montreal Neurological Institute (MNI) template space, and finally spatially smoothed the images using an 8-mm FWHM Gaussian kernel. Data were filtered using a high-pass filter with 1/128 Hz cutoff frequency.

General linear model (GLM) analyses

We constructed the following GLMs to address specific questions. We first focused our analyses of inequality processing on striatum and ventromedial prefrontal cortex (VMPFC) in GLM 1, since these regions have been repeatedly suggested to be critically involved in equality processing (3). We thus performed region-of-interest (ROI) analyses of the parametric effect of equality within the meta-analytic functional coactivation masks for “Striatum” and “VMPFC” in the Neurosynth database (https://neurosynth.org/), which were independent of our specific GLM contrasts. We complemented these ROI analyses with exploratory whole-brain analyses to identify any other areas responding to inequality. With GLM 1, we also explored neural responses to harm signals with whole-brain analyses. Second, we constructed GLM 2 in which we examined neural representations of integrated utility, which was derived from the computational behavioral model and the equality and harm magnitudes in each trial. Thus, we applied a VMPFC region (peak MNI coordinates [0, 52, -8]) which was suggested as the peak voxel involved in monetary incentives processing in a meta-analyses study (4) as a ROI to test the neural validation of our computational behavioral model (i.e., whether activity in VMPFC represents the model-predicted value of the chosen option). Third, with GLM 3 and 4, we further performed psychophysiological interaction (PPI) analyses to clarify potential neural networks underlying different motives and observed behaviors. In GLM 3, we split trials with different levels of equality signals into two bins (i.e., high vs low equality signals) for easy-to-visualize PPI analyses examining how interregional functional connectivity (i.e. connectivity between Striatum and other regions) varies with equality levels. In GLM 4, we contrasted choices of the more unequal vs the more equal offer in the Rank-reversal condition, and further tested how functional connectivity between Striatum and other regions varies with different choices and strengths of different motives. For any question for which we did not have *a priori* hypotheses - such as contrasts between different types of choices, and PPI analyses – we conducted whole-brain analyses to identify the relevant areas, followed by more sensitive ROI post-hoc tests controlling for specific potential confounds, with ROIs defined as spheres with 8 mm radius around the center MNI coordinates of the respective regions.

Specifically, we built GLM 1 to examine how signals of equality and harm to others were represented in the brain during wealth redistribution in different conditions. GLM 1 included the regressor corresponding to the onset of alternative offer presentation in each condition separately (i.e., No Rank-reversal, Rank-reversal, and filler). The duration for these events were equal to the time form onsets of the alternative offer presentation to the time points of offsets. Moreover, GLM 1 included the trial-wise equality difference between the two alternative offers ($-\Delta F$) as the parametric modulator for the alternative offer events in the No Rank-reversal condition, and both the trial-wise equality difference between the two alternative offers ($-\Delta F$) and trial-wise harm of the more equal alternative offer (H) as the parametric modulators for the alternative offer events in the Rank-reversal condition. To identify neural correlates that reflected the signals of equality and harm irrespective of participants' choices, we examined the following contrasts: 'No Rank-reversal: equality signal (-ΔF)', 'Rank-reversal: equality signal ($-\Delta F$)', and 'Rank-reversal: harm signal (H)', respectively. To identify neural correlates that reflected the difference in the signal of equality difference between the No Rank-reversal and Rank-reversal condition, we examined the contrast of 'No Rank-reversal:

equality signal > Rank-reversal: equality signal'. Significant results are reported at a cluster-wise FWE corrected $p < 0.05$ (cluster-forming threshold voxel- wise $p < 0.001$ uncorrected) throughout all the analyses unless otherwise noted.

The correlations (i.e., Pearson r) between the equality and harm signals ranged from 0.45 to 0.59. To ensure that the multiple parametric modulators included in the GLM will not introduce a collinearity problem, we followed established procedures (5) to perform collinearity diagnostics analyses. The results showed that the condition index values for the parametric modulators ranged between 1.00 and 1.58, and thus much lower than the tolerance threshold of 30, suggesting that this GLM was not likely degraded by the presence of collinearity. Note that we did not differentiate neural signals of rank reversal in this GLM, since rank reversal was manipulated by including two different conditions - No Rank-reversal and Rank-reversal – that also differ in other variables, such as the amount of transferred money and response bias.

To confirm that the weaker parametric effect of equality signals in the striatum for the Rank-reversal condition was not because the parametric modulator of harm signals in the Rank-reversal condition captured the variance accounted by equality signals, we established GLM 1a in which only trial-wise equality signal ($-\Delta F$) was included as the parametric modulator for the alternative offer events in both conditions. Therefore, in the GLM 1a, the alternative offer onsets in the No Rank-reversal and Rank-reversal condition have the same parametric modulator (i.e., equality signal ($-\Delta F$)). This revealed the same results as GLM 1 (Table S8, Figure S6).

To visualize the neural response patterns of the parametric effects identified in GLM 1, we generated GLM 1b in which trials were divided into 4 levels of equality signals (i.e., -8, -6, -4, and -2) and 4 different alternative offer onset regressors corresponding to the 4 equality levels were included for each condition (i.e., No Rank-reversal and Rank-reversal), which resulted in 8 regressors of interest. By depicting the betas from GLM 1b, we can thus visualize how neural responses are modulated by equality signals and how such neural equality signals differ across different conditions (Figure S7).

In GLM 2, we tested regions computing decision utility (i.e., utility of the chosen offer) which was defined based on the winning model in the Rank-reversal condition. To this end, we included the regressor corresponding to the onset of alternative offer presentation in each condition separately (i.e., No Rank-reversal, Rank-reversal, and filler) in the same way as GLM 1. Since we only applied computational modeling analyses in the Rank-reversal condition, we only included the trial-wise utility of the chosen offer and the utility of the unchosen offer as the parametric modulators for the alternative offer events in the Rank-reversal condition. Since the decision utility was derived from the winning model and the same equality and harm signals used in GLM 1, including utility of the chosen and unchosen offer together with the equality and harm signals within a single GLM would generate multiple collinearity problem. Therefore, we constructed GLM 2 separately from GLM 1. The durations for these events were equal to the time form onsets of the alternative offer presentation to the time points of offsets. The collinearity diagnostics analyses showed that the condition index values ranged between 1.00 and 4.76, which were much lower than the tolerance threshold of 30, suggesting that this GLM was not likely degraded by the presence of collinearity.

To further investigate how striatum changed its functional coupling with other brain regions as a function of equality signals, we constructed GLM 3. For ease of visualization and interpretation of the PPI results, we divided trials in this GLM into 2 levels of equality difference between alternative offers (i.e. high equality difference: $-\Delta F$ = -2 and -4, vs low equality difference: ΔF = -6 and -8) and 2 different alternative offer onset regressors corresponding to the 2 equality levels were included for each condition (i.e., No Rank-reversal and Rank-reversal). In line with GLM 1 results, GLM 3 consistently suggested that in the No Rank-reversal condition, activity in the striatum was stronger for higher equality difference than lower equality difference, and the effect was not observed in the Rank-reversal condition. Importantly, we performed easy-to-visualize functional connectivity analyses based on GLM 3 to examine how the striatum connects with other brain regions depending on different equality levels and different conditions.

In GLM 4, we identified neural activity associated with specific choices during wealth redistribution in both conditions. Thus, GLM 4 included onsets of alternative offer presentation of each condition with respect to specific choice (i.e., equal choice or unequal choice in the No Rank-reversal and Rank-reversal condition), resulting in four regressors of interest. To control for any potential effect of utility of each offer on trials with regard to specific choice in the Rank-reversal condition, the trial-wise utility of the chosen and the unchosen options were included as parametric modulators for each choice onset regressor. We also constructed GLM 4a in which the utility of the chosen and unchosen options were not included as parametric modulators (Table S11). For GLM 4 modeling responses in both conditions, we excluded 11 participants who always chose one type of choice in either condition; and for analyses only involved in the Rank-reversal condition, we excluded 7 participants who always chose one type of choice in the Rank-reversal condition.

For the GLMs above, all parametric regressors were z-standardized before being entered into the GLM analyses. We switched off orthogonalization during model estimation to allow the parametric modulators to compete for variance. For all the GLMs, we incorporated onsets of fixation, initial offer presentation, and alternative offer presentation of trials with no response as regressors of no interest. For GLM 1, GLM 1a, and GLM 2, trail-wise difference in initial payoff between the two parties ($\Delta\ Initial\ endowment$) was included as the parametric modulator for the initial offer event onset. In addition, we included six rigid body parameters as regressors of no interest, to account for head motion artifacts. Regressors of interest and no interest were convolved with a canonical hemodynamics response function (HRF). For the GLMs, the cosine values (calculated by SPM) between different regressors ranged between -0.5 and 0.5 (colinear if consine = +1/-1), also indicating that these GLMs did not likely suffer from collinearity problems.

We fed individual-level contrasts into group-level random-effect analyses with one-sample t tests to assess the neural parametric effects of signals for inequality, harm to others, and decision utility, or to compare parametric contrasts between the No Rank-reversal and Rank-reversal condition. Flexible factorial analyses were used to examine potential interaction effects. Correlation analyses were used to explore potential relationship between different prosocial motives (i.e., α, β, and δ) and neural activities. Since the distributions of α and β were positively skewed (skewness(α) = 0.49 and skewness(β) = 1.64), we normalized these two parameters by taking the square roots of the parameters, which were more normally distributed (skewness ($\alpha - \mathrm{normalized}$) = - 0.11, and skewness ($\beta - \mathrm{normalized}$) = 0.25), as the measures of inequality aversion and harm aversion in correlation analyses.

Inference for all whole-brain GLM and PPI analyses was corrected for multiple comparisons across the whole brain at a cluster-level threshold of $p < 0.05$ family-wise error (FWE) corrected, with an initial cluster-forming height threshold of $p < 0.001$ uncorrected. All whole-brain analyses employed non-parametric tests (5000 permutations) implemented in the SnPM package (https://warwick.ac.uk/snpm), which minimized the risk of type-1 error rate inflation (6).

Inference for all ROI analyses was performed at voxel-level $p < 0.05$ family-wise error (FWE) correction for the ROI volume. The “Striatum” and “VMPFC” masks for ROI analyses were defined based on meta-analytic functional coactivation maps of “Striatum” and “VMPFC” in the Neurosynth database (https://neurosynth.org/), and the peak MNI coordinates of the Striatum [-12, 10, -6] were also derived from this activation map for post-hoc tests. For post-hoc tests of specific contrasts that controlled for potential confounds, the small volumes were defined as spheres with 8 mm radius around center MNI coordinates of the whole-brain-corrected contrasts. Specific center MNI coordinates for different analyses are reported in the Results section.

Functional connectivity analyses

By performing functional connectivity analyses, we aimed to address two questions: 1) whether the neural sensitivity to equality in the striatum in the Rank-reversal condition was modulated by other cognitive processes, especially when harm aversion/rank reversal aversion conflict with inequality aversion; 2) whether different motives (i.e., harm aversion and rank reversal aversion) interact with the striatum through different systems to affect redistribution decisions. To address these questions, we followed (7) to establish two PPI models. To answer the first question, we took the striatum (i.e., a 6-mm radius sphere region centered at the peak MNI coordinates of [-12, 10, -6] of the meta-analytical "striatum" mask from Neurosynth) as the seed region in the PPI analyses. We conducted a PPI analysis for each of the two conditions (i.e., No Rank-reversal and Rank-reversal) to assess differential functional connectivity with this seed in high $- \Delta F$ trials compared with low $- \Delta F$ trials based on GLM 3. For each PPI analysis, the BOLD signal within the seed (i.e., average time series within 6-mm sphere around the peak voxel) was used as the physiological factor and high equality signals $(- \Delta F)$ versus low equality signals $(- \Delta F)$ contrast in GLM 3 was used as the psychological factor. At the first level, the PPI model included one regressor representing the extracted time series in the seed (i.e., the physiological variable), one regressor representing the psychological variable of interest, and a third regressor representing the interaction of the two regressors (the PPI term). Note that the interaction term therefore identifies areas that show context-dependent connectivity while fully controlling for any simple effect of the seed time course and the psychological factor.

To answer the second question, we took the striatum region that was associated with equality processing and equal choice (peak MNI coordinates: [-18, 11, -2]) identified in GLM 1 as the seed region. Since for these PPI analyses, we aimed to examine the neural network underlying redistribution decisions and how different motives affected decisions when there were conflicts between motives, we focused on the contrast of "more unequal choice > more equal choice" for the Rank-reversal condition in GLM 4. Therefore, the PPI models included one regressor representing the extracted time series in the seed (a 6-mm sphere region centered at coordinates corresponding to the striatum) as the physiological variable, one regressor representing the psychological variable of interest (i.e., more unequal choice > more equal choice), and a third regressor representing the interaction of the two regressors (the PPI term).

At the second level, for the first PPI analyses, two beta maps of the PPI term in the No Rank-reversal and Rank-reversal conditions for each participant were fed into a paired t-test analyses. For the second PPI analyses, since we were interested in how different motives modulate the neural networks to affect decisions, we correlated the beta maps with individuals' parameters of inequality aversion (α), harm aversion (β), or rank reversal aversion (δ) derived from the winning computational model. Significant results were reported with a whole-brain corrected threshold [i.e., a combined threshold of voxel-level $p < 0.001$ uncorrected and cluster-level $p < 0.05$ family-wise error (FWE) correction] unless a special note.

SI Results

In the analyses of GLM 4, we showed that greater activity in DMPFC and TPJ were associated with stronger inequality aversion, and that greater activity in putamen was associated with stronger harm aversion when people choosing the more unequal offer. These correlation patterns still held after controlling for the effect of the other two parameters (for DMPFC – α [β and δ controlled]: peak MNI coordinates: [-3, 53, 31], t-value = 2.88, voxel-wise *p* (FWE-SVC) = 0.040, k =3, ROI center MNI coordinates [0, 59, 28]; for TPJ – α [β and δ controlled]: peak MNI coordinates: [-57, -64, 25], t-value = 3.46, voxel-wise *p* (FWE-SVC) = 0.011, k =26, ROI center MNI coordinates [-54, -61, 25]; for Putamen – β [α and δ controlled]: peak MNI coordinates: [-21, -4, -11], t-value = 3.56, voxel-wise *p* (FWE-SVC) = 0.009, k =50, ROI center MNI coordinates [-24, -1, -5]).

In the whole-brain PPI analyses of GLM 3, we showed that DMPFC (peak MNI coordinates: [0, 47, 40]) was functionally connected with striatum (center MNI coordinates [-12, 10, -6]) more strongly for higher equality signals (i.e., high - ΔF vs low - ΔF) in the Rank-reversal condition. To confirm this effect in parametric analysis, we also performed a post-hoc PPI analyses based on GLM 1 (parametric analyses) with the same seed striatum region (center MNI coordinates [-12, 10, -6]), and took the DMPFC region (peak MNI coordinates: [0, 47, 40]) identified in the above analysis as a ROI. This analysis revealed convergent effect that the DMPFC–Striatum connectivity was stronger for higher equality signals (higher - ΔF) in the Rank-reversal condition (peak MNI coordinates: [-3, 44, 43], t-value = 3.22, voxel-wise *p* (FWE-SVC) = 0.022, k =67, ROI center MNI coordinates [0, 47, 40]).

In the PPI analyses of GLM 4, we revealed association between stronger Striatum-IFG connectivity and greater inequality aversion, and association between stronger Striatum-SFG connectivity and greater rank reversal aversion. These correlation patterns of different networks also held after controlling for the effect of the other two parameters (for Striatum-IFG with α [β and δ controlled]: peak MNI coordinates: [57, 23, 7], t-value = 4.44, voxel-wise *p* (FWE-SVC) = 0.001, k =35, ROI center MNI coordinates [57, 23, 13]; for Striatum-SFG with δ [α and β controlled]: peak MNI coordinates: [-21, 2, 49], t-value = 3.41, voxel-wise *p* (FWE-SVC) = 0.007, k =42, ROI center MNI coordinates [-24, -1, 49]).

References

1. W. Xie, B. Ho, S. Meier, X. Zhou, Rank reversal aversion inhibits redistribution across societies. *Nat. Hum. Behav.* **1**, 0142 (2017).
2. S. Lewandowsky, S. Farrell, Computational Modeling in Cognition: Principles and Practice (2011) https:/doi.org/10.4135/9781483349428.
3. E. Tricomi, A. Rangel, C. F. Camerer, J. P. O'Doherty, Neural evidence for inequality-averse social preferences. *Nature* **463**, 1089–91 (2010).
4. O. Bartra, J. T. McGuire, J. W. Kable, The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).
5. D. Belsley, E. Kuh, R. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (John Wiley & Sons, Inc., 1980).
6. T. E. Nichols, A. P. Holmes, Nonparametric Permutation Tests For Functional Neuroimaging : A Primer with Examples. **25**, 1–25 (2001).
7. K. J. Friston, *et al.*, Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218–229 (1997).

Table S1. Mixed-effects model results of behavioral data in fMRI wealth redistribution task

	Generalized mixed-effects model		
Variable	*B (S.E.)*	*ORE (95% CI)*	*P Value*
Intercept	1.15(0.18)	3.17(2.21, 4.53)	< 0.001
Condition (C)	-0.77 (0.04)	0.37 (0.33, 0.42)	< 0.001
∆ Inequality (I)	0.46 (0.07)	1.58 (1.37, 1.83)	< 0.001
∆ Initial Endowment (E)	0.11 (0.05)	1.12 (1.01, 1.24)	0.04
∆ Transfer (T)	-0.77 (0.04)	0.46 (0.43, 0.50)	< 0.001
C*I	-0.14 (0.09)	0.87 (0.73, 1.03)	0.11
C*E	-0.02 (0.05)	0.98 (0.88, 1.09)	0.70
I*E	-0.06 (0.07)	0.94 (0.82, 1.09)	0.44
I*T	-0.37 (0.16)	0.69 (0.50, 0.96)	0.03
E*T	0.17 (0.04)	1.19 (1.10, 1.28)	< 0.001
C*I*E	-0.02 (0.09)	0.98 (0.82, 1.16)	0.79
C*I*T	0.37 (0.11)	1.44 (1.16, 1.79)	< 0.001
I*E*T	0.11(0.16)	1.12 (0.82, 1.53)	0.48
C*I*E*T	-0.11 (0.11)	0.90 (0.72, 1.11)	0.31
LL		-4642	
BIC		9422	
Marginal R^2		0.19	

ORE, odds ratio estimate; CI, confidence interval; LL, log-likelihood; BIC, Bayesian Information Criterion

Table S2. Main effect of Δ Inequality on probability to choose the more equal alternative offer

Δ Inequality	P (Equal choice) (Mean ± SE)
2	0.52 ± 0.02
4	0.58 ± 0.03
6	0.62 ± 0.03
8	0.61 ± 0.03

Table S3. Main effect of ∆ Initial endowment on probability to choose the more equal alternative offer

∆ Initial endowment	P (Equal choice) (Mean ± SE)
Low	0.57 ± 0.03
High	0.59 ± 0.03

Table S4. Main effect of ∆ Transfer on probability to choose the more equal alternative offer

∆ Transfer	P (Equal choice) (Mean ± SE)
1 - 3	0.76 ± 0.03
4 - 6	0.522 ± 0.03
7 - 11	0.36 ± 0.04

Table S5. Interaction between Δ Inequality and Δ Transfer on probability to choose the more equal alternative offer in the Rank-reversal condition

	Δ Transfer		
	Low (3-5)	Middle (6-8)	High (9-11)
Δ Inequality	(Mean ± SE)	(Mean ± SE)	(Mean ± SE)
2	0.27 ± 0.04	0.28 ± 0.05	0.27 ± 0.04
4	0.38 ± 0.05	0.39 ± 0.05	0.36 ± 0.05
6	0.46± 0.05	0.42 ± 0.05	0.45 ± 0.05
8	0.51 ± 0.05	0.40 ± 0.05	0.41 ± 0.05

Table S6. Interaction between Δ Initial Endowment and Δ Transfer on probability to choose the more equal alternative offer in the Rank-reversal condition

	Δ Transfer		
Δ Initial Endowment	Low (3-5) (Mean ± SE)	Middle (6-8) (Mean ± SE)	High (9-11) (Mean ± SE)
Low	0.76 ± 0.03	0.49 ± 0.03	0.34 ± 0.04
High	0.75 ± 0.02	0.59 ± 0.03	0.38 ± 0.04

Table S7. Results of whole-brain parametric analysis of fMRI data in GLM 1

		Peak MNI coordinates				
Regions	Laterality	x	y	z	t-value	Extent (k)
No Rank-reversal (whole-brain): Positive association with $-\Delta F$ (critical extent k > 106 voxels)						
Caudate/Putamen	L	-15	5	19	4.33	364
	R	12	20	-5	4.00	127
IOG	R	39	-88	-2	4.70	1546
MOG	L	-27	-93	4	4.53	290
Rank-reversal (whole-brain): Positive association with $-\Delta F$ (critical extent k > 104 voxels)						
No significant cluster						
Rank-reversal (whole-brain): Positive association with harm to others (H) (critical extent k > 115 voxels)						
TPJ	L	-42	-46	46	7.74	4116
	R	39	-49	43	7.95	
DMPFC/ACC	L	-3	29	37	6.45	717
IFG	R	45	14	31	7.35	1421
MFG	L	-48	14	37	6.02	1019
ITG	R	54	-49	-23	4.92	327

IOG, inferior occipital gyrus; MOG, middle occipital gyrus; TPJ, temporoparietal junction; DMPFC, dorsomedial prefrontal cortex; ACC, anterior cingulate cortex; IFG, inferior frontal gyrus; MFG, middle frontal gyrus; ITG, inferior temporal gyrus. ΔF, inequality difference between alternative offers; H, harm amount, the proportion of the transferred money that exceeds the necessary amount of money that can reach the same equality level but not reverse the initial relative rankings for the more equal alternative offer. For whole-brain analyses, significant clusters were thresholded at voxel-wise $p < 0.001$ uncorrected and cluster-wise FWE corrected $p < 0.05$. Extent threshold for each contrast was determined by the permutation test implemented in SnPM.

Table S8. Results of whole-brain/ROI parametric analysis of fMRI data in GLM 1a

		Peak MNI coordinates				
Regions	Laterality	x	y	z	t-value	Extent (k)
No Rank-reversal (whole-brain): Positive association with $-\Delta F$ (critical extent k > 108 voxels)						
Caudate/Putamen	L	-15	5	19	4.37	373
	R	12	20	-5	4.08	140
MOG	L	-27	-94	4	4.59	304
	R	30	-85	22	4.78	1591
Rank-reversal (whole-brain): Positive association with $-\Delta F$ (critical extent k > 94 voxels)						
No significant cluster						
No Rank-reversal ("Striatum" mask): Positive association with $-\Delta F$						
Caudate/Putamen	L	-18	11	1	3.98	120
	R	15	20	-5	3.82	90
No Rank-reversal > Rank-reversal ("Striatum" mask): Association with $-\Delta F$						
Caudate	L	-18	17	2	3.44	7
	R	6	14	-5	3.39	18

MOG, middle occipital gyrus. ΔF, inequality difference between alternative offers. For whole-brain analyses, significant clusters were thresholded at voxel-wise $p < 0.001$ uncorrected and cluster-wise FWE corrected $p < 0.05$. Extent threshold for each contrast was determined by the permutation test implemented in SnPM. For ROI analyses, significant clusters were thresholded at voxel-wise FWE corrected $p < 0.05$.

Table S9. Results of whole-brain analysis of fMRI data in GLM 4

		Peak MNI coordinates				
Regions	Laterality	x	y	z	t-value	Extent (k)
No Rank-reversal: unequal choice > equal choice (critical extent k > 86 voxels)						
MFG/IFG	R	51	20	16	5.39	1234
IFG/Insula	L	-30	26	-5	5.90	1210
ACC	L, R	0	35	25	4.83	926
MCC	R	3	-37	40	4.23	113
TPJ	R	54	-49	34	4.72	272
IPL	L	-45	-58	46	4.03	175
No Rank-reversal: equal choice > unequal choice (no voxel with p < 0.001 uncorrected)						
No significant cluster						
Rank-reversal: unequal choice > equal choice (no voxel with p < 0.001 uncorrected)						
No significant cluster						
Rank-reversal: equal choice > unequal choice (critical extent k > 22 voxels)						
No significant cluster						
Rank-reversal: unequal choice > equal choice, positively correlated with inequality aversion (α) (critical extent k > 46 voxels)						
TPJ	L	-54	-61	25	4.27	304
DMPFC	L	0	59	28	4.48	167
Rank-reversal: unequal choice > equal choice, positively correlated with harm aversion (β) (critical extent k > 81 voxels)						
Putamen	L	-24	-1	-5	3.63	94
MOG	R	24	-94	4	4.18	309
IOG	L	-51	-64	-17	3.87	520

MFG, middle frontal gyrus; IFG, inferior frontal gyrus; ACC, anterior cingulate cortex; MCC, middle cingulate cortex; TPJ, temporoparietal junction; IPL, inferior parietal lobe; DMPFC, dorsomedial prefrontal cortex; MOG, middle occipital gyrus; IOG, inferior occipital gyrus. Significant clusters were thresholded at voxel- wise p < 0.001 uncorrected and cluster-wise FWE corrected p < 0.05. Extent threshold for each contrast was determined by the permutation test implemented in SnPM.

Table S10. Results of post-hoc ROI interaction analysis of fMRI data in GLM 4

		Peak MNI coordinates					
Regions	Laterality	x	y	z	t-value	voxel-wise *p*	Extent (k)
Interaction: No Rank-reversal (unequal choice > equal choice) > Rank-reversal (unequal choice > equal choice)							
MFG/IFG	R	48	17	13	2.91	0.025	22
IFG/Insula	L	-27	29	-5	2.90	0.026	12
ACC	L, R	-6	35	28	3.16	0.014	38
MCC	R	0	-37	40	3.38	0.006	56
TPJ	R	48	-46	34	3.68	0.002	77
IPL	L	-48	-52	43	2.85	0.029	15

MFG, middle frontal gyrus; IFG, inferior frontal gyrus; ACC, anterior cingulate cortex; MCC, middle cingulate cortex; TPJ, temporoparietal junction; IPL, inferior parietal lobe. Significant clusters were thresholded at voxel-wise FWE corrected $p < 0.05$. The ROI center MNI coordinates were selected based on the whole-brain analyses reported in Table S9.

Table S11. Results of whole-brain analysis of fMRI data in GLM 4a

		Peak MNI coordinates				
Regions	Laterality	x	y	z	t-value	Extent (k)
No Rank-reversal: unequal choice > equal choice (critical extent k > 89 voxels)						
MFG/IFG	R	51	20	16	5.29	1228
IFG/Insula	L	-30	26	-5	5.80	1132
ACC	L, R	0	35	25	4.75	894
MCC	R	3	-37	40	4.29	121
TPJ	R	54	-49	34	4.77	283
IPL	L	-45	-58	46	4.12	183
No Rank-reversal: equal choice > unequal choice (no voxel with *p* < 0.001 uncorrected)						
No significant cluster						
Rank-reversal: unequal choice > equal choice (critical extent k > 88 voxels)						
MOG	R	30	-82	34	3.99	115
Fusiform gyrus	L	-36	-61	-17	3.80	116
	R	33	-61	-17	3.78	169
IPL	L	-30	-46	55	3.76	116
Lingual gyrus	L	-15	-97	-14	3.70	98
Rank-reversal: equal choice > unequal choice (no voxel with *p* < 0.001 uncorrected)						
No significant cluster						

MFG, middle frontal gyrus; IFG, inferior frontal gyrus; ACC, anterior cingulate cortex; MCC, middle cingulate cortex; TPJ, temporoparietal junction; IPL, inferior parietal lobe. Significant clusters were thresholded at voxel-wise *p* < 0.001 uncorrected and cluster-wise FWE corrected *p* < 0.05. Extent threshold for each contrast was determined by the permutation test implemented in SnPM.

Table S12. Results of whole-brain PPI analysis of unequal choice vs equal choice in the Rank-reversal condition

		Peak MNI coordinates				
Regions	Laterality	x	y	z	t-value	Extent (k)
PPI: unequal choice > equal choice, seed Striatum centered at [-18, 11, -2]						
Positively correlated with inequality aversion (α) (critical extent k > 97 voxels)						
IFG	R	57	23	13	5.08	120
Positively correlated with rank reversal aversion (δ) (critical extent k > 106 voxels)						
SFG	L	-24	-1	49	5.35	145

IFG, inferior frontal gyrus; SFG, superior frontal gyrus. Significant clusters were thresholded at voxel-wise $p < 0.001$ uncorrected and cluster-wise FWE corrected $p < 0.05$. Extent threshold for each contrast was determined by the permutation test implemented in SnPM.

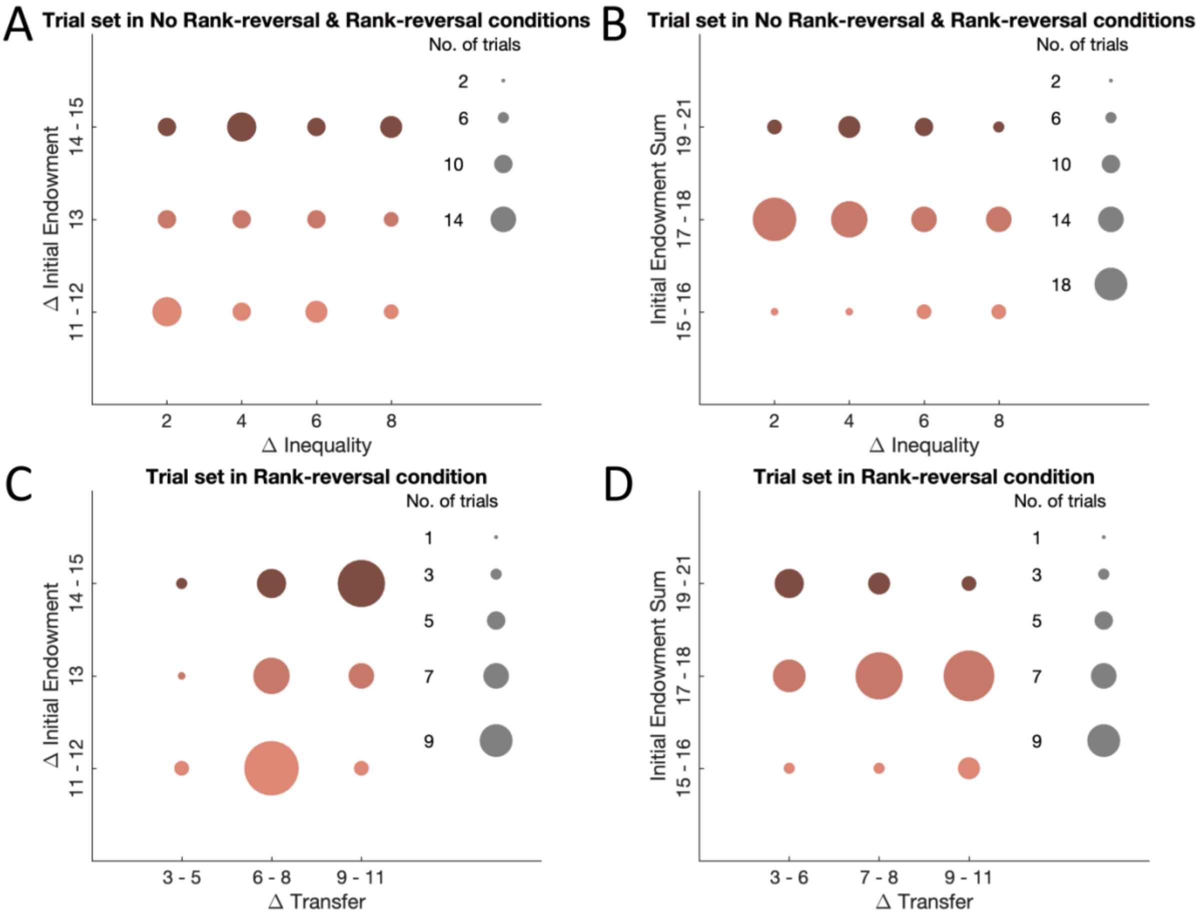


Figure S1. Trial set matrix of the No Rank-reversal and Rank-reversal conditions. To clarify the effects of inequality and the amount of transferred money from the advantaged party, we orthogonalized the differences in inequality/transferred money between the two alternative offers with the difference/sum of the initial endowment of the two parties. (A & B) Design matrix with x axis representing the difference in inequality level (Δ Inequality) between the two alternative offers, and y axis representing the difference between the two parties in initial endowment (A) and the sum of initial endowment for the two parties (B). (C & D) Design matrix with x axis representing the difference in transferred money (Δ Transfer) between the two alternative offers, and y axis representing the difference between the two parties in initial endowment (C) and the sum of initial endowment for the two parties (D). The size of the circle is proportional to the number of trials in each type of variable combination.

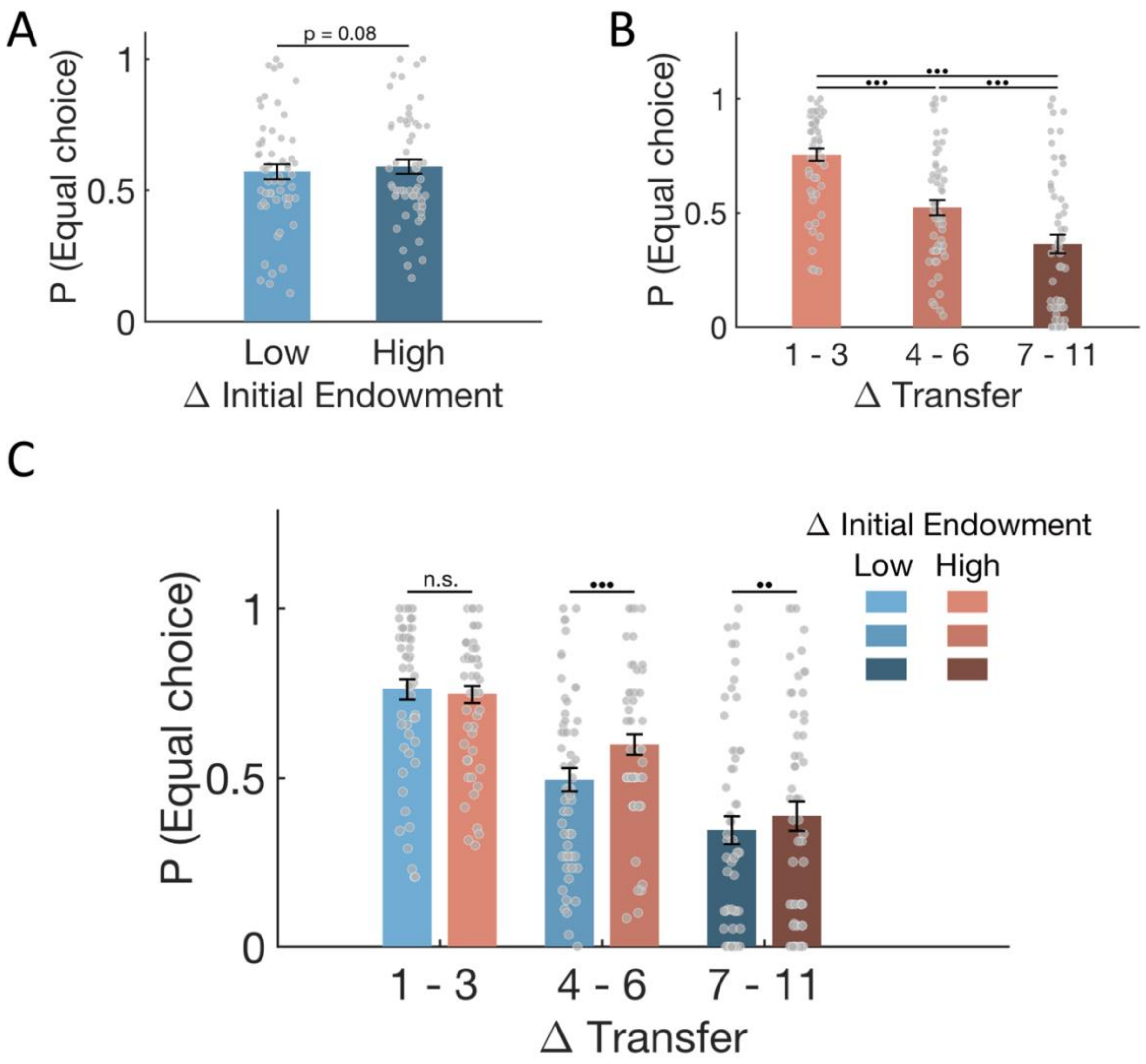


Figure S2. Behavioral results. (A) Main effect of difference in initial endowment (Δ Initial endowment) between the two parties on probability to choose the more equal offer. Participants chose the more equal offer more frequently when initial endowment difference was higher. **(B)** Main effect of difference in transferred money (Δ Transfer) between the two alternative offers on probability to choose the more equal offer. Probability to choose the more equal offer decreased with the increase of difference in transferred money between the more equal offer and the more unequal offer. **(C)** Interaction between Δ Initial endowment and Δ Transfer on probability to choose the more equal offer. When the difference in transferred money is high (i.e., 4 - 6 and 7 - 11), higher initial endowment difference increased the probability to choose the more equal offer. Each grey dot represents one participant, and error bars represent the SEMs. •••, $p < 0.001$; ••, $p < 0.01$; •, $p < 0.05$, n.s., $p > 0.1$.

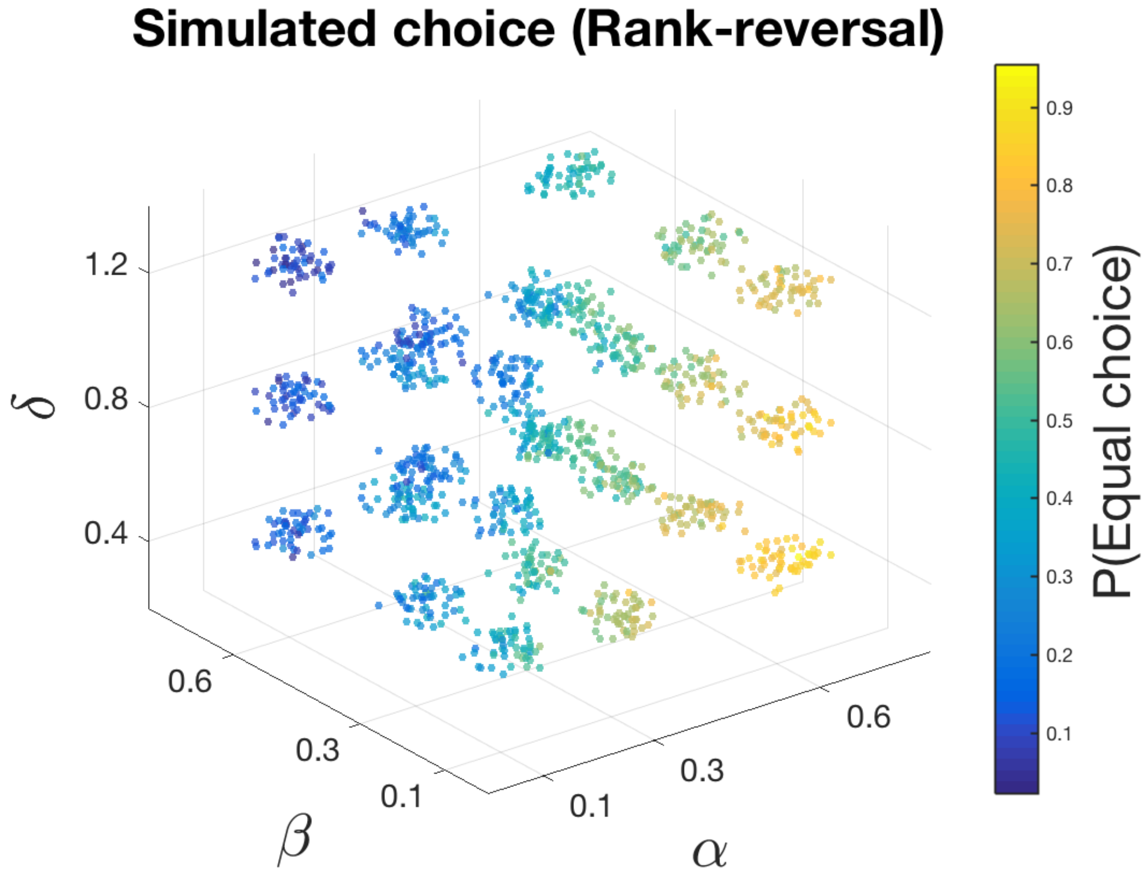


Figure S3. Simulation results of the winning model. The 3-D scatter plot shows that the simulated responses [i.e., P(Equal choice)] vary with the three parameters, indicating that different motives affect individuals' redistribution decisions in different manners. To do this analysis, we generated 27 datasets using all combinations of three plausible values for each parameter (α: 0.1, 0.3, 0.6; β: 0.1, 0.3, 0.6; δ: 0.8, 1.1, 1.4). The temperature parameter λ was fixed at 1.2. Each dot represents one repeat of simulation. For each repeat, a small noise derived from a uniform distribution (0, 0.1) was added to each of the three parameter values in a certain parameter combination. We simulated responses with the winning model in the Rank-reversal condition 50 times for each parameter combination. The color of each point indicates the probability of more equal choice in each simulation.

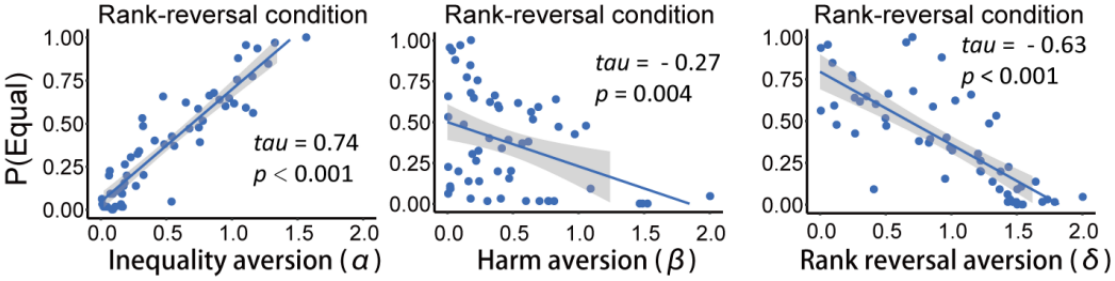


Figure S4. Correlations between probability to choose the more equal offer and model parameters which correspond to different motives in the Rank-reversal condition. Scatter plots show that individuals who are more averse to inequality (i.e., higher α, left panel) will choose the more equal offer more frequently, and individuals who are more averse to harming others (i.e., higher β, middle panel) and rank reversal (i.e., higher δ, right panel) will choose the more equal offer less frequently. Each dot represents one participant.

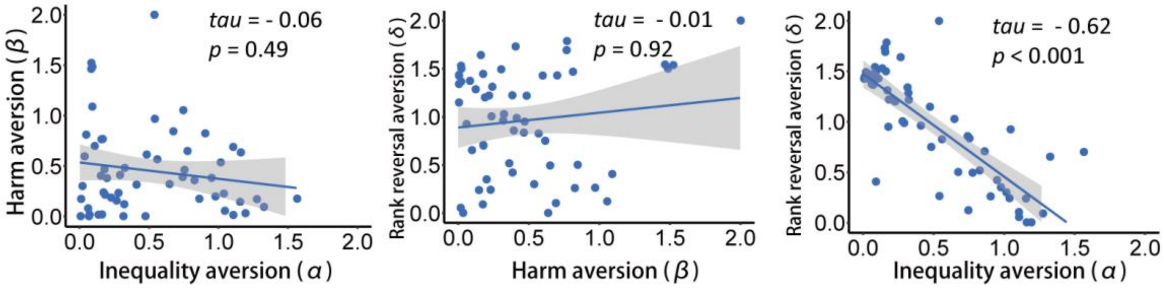


Figure S5. Correlation analyses between the three model parameters in the winning model (M4a). Scatter plots show that α (inequality aversion) and δ (rank reversal aversion) are negatively associated with each other. Each dot represents one participant.

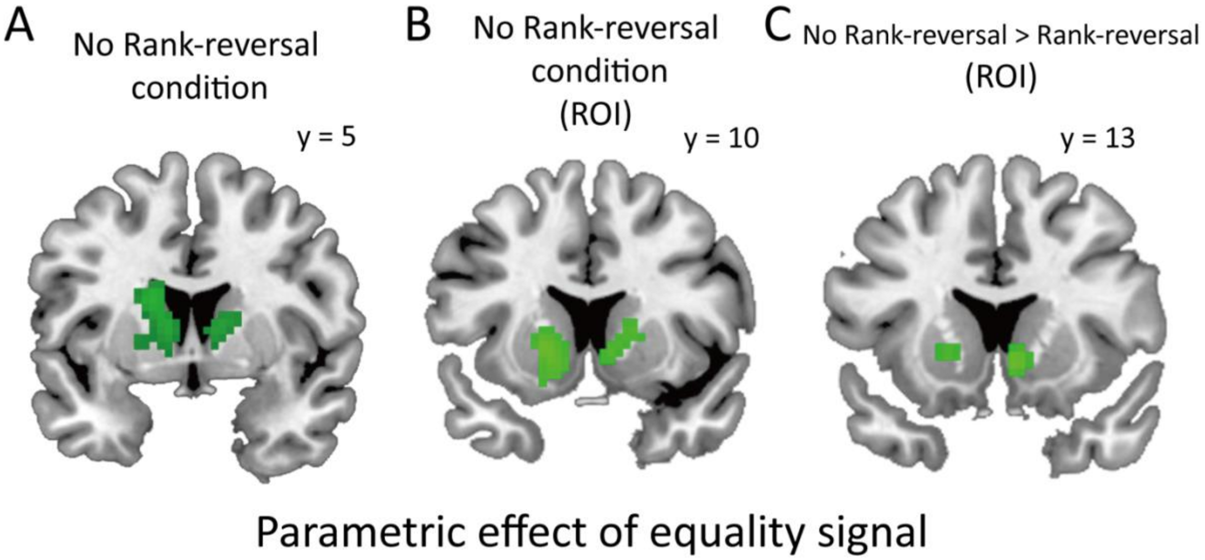

A
No Rank-reversal condition
y = 5
B
No Rank-reversal condition (ROI)
y = 10
C
No Rank-reversal > Rank-reversal (ROI)
y = 13
Parametric effect of equality signal

Figure S6. Results of parametric analyses in GLM 1a. Activity in caudate/putamen was associated with the equality signal (-ΔF) in the No Rank-reversal condition for both whole-brain **(A)** and independent ROI analyses (with the "Striatum" mask) **(B)**. **(C)** ROI analysis further shows a greater parametric strength of equality in striatum in the No Rank-reversal condition than in the Rank-reversal condition. Results of GLM 1a show similar results as GLM 1, confirming the parametric effects of ΔF identified in GLM 1.

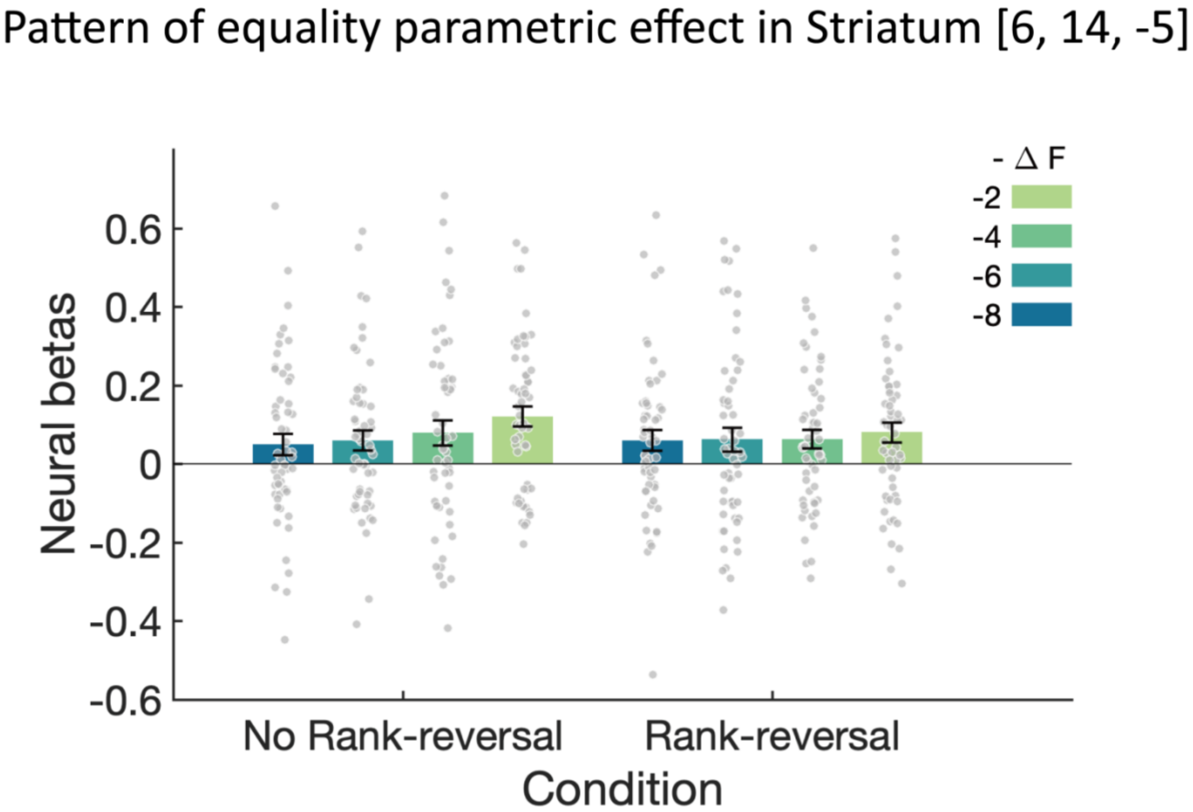


Figure S7. Response pattern for parametric effects of equality in striatum identified in GLM 1. To visualize the effect of parametric modulation of inequality difference identified in GLM 1, we separately regressed the four levels of equality difference (i.e., -ΔF = -2, -4, -6, and -8) in each condition, and constructed and re-estimated GLM 1b and extracted neural betas in the significant cluster (i.e., caudate/putamen) identified in GLM 1. In line with GLM 1 results, activity in striatum increased with the increase of equality signal (-ΔF) in the No Rank-reversal condition, but no such effect was observed in the Rank-reversal condition. To avoid double dipping the data, we did not perform any statistical analyses here. Each dot represents one participant, and error bars represent the SEMs.

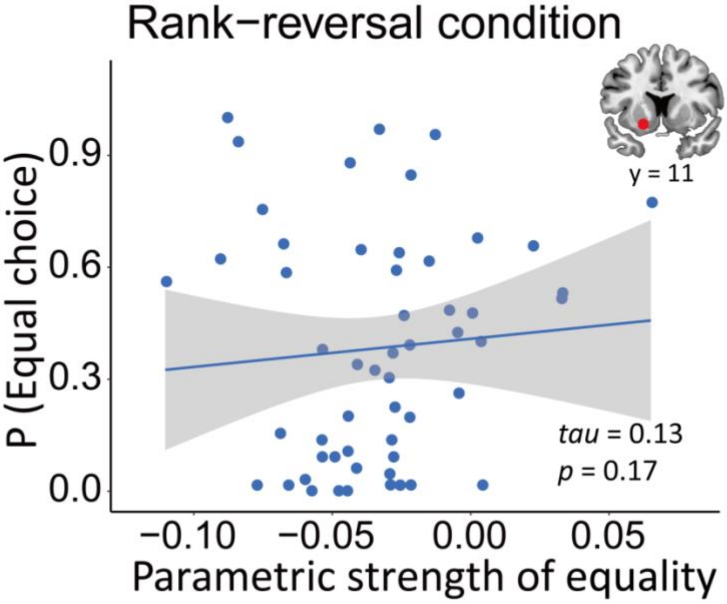
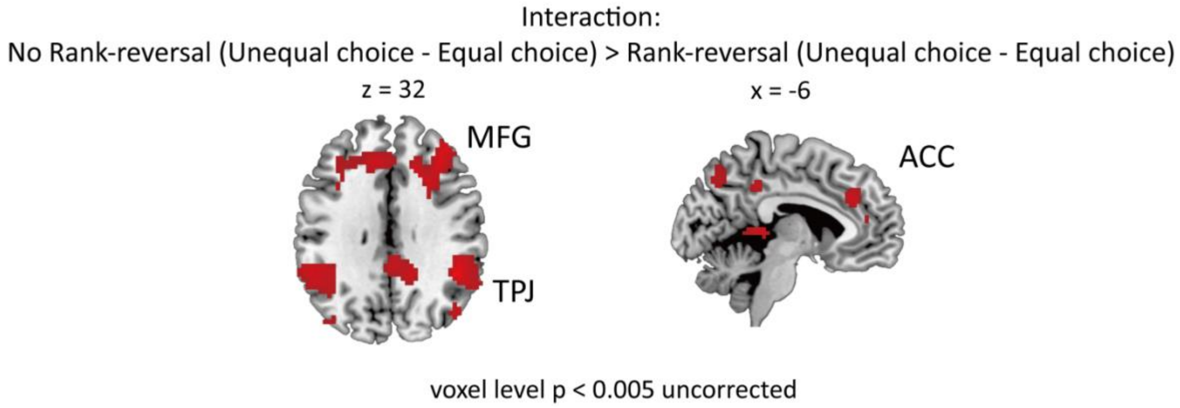


Figure S8. Correlation between the parametric strength of equality in striatum and the probability to choose the more equal offer in the Rank-reversal condition. Scatter plot shows that the correlation between the parametric strength of equality in striatum (a region with the center of peak MNI coordinates: [-12, 10, -6] in the "Striatum" mask from Neurosynth) and individuals' probability to choose the more equal offer in the Rank-reversal condition is not significant. Each dot represents one participant.



Interaction:
No Rank-reversal (Unequal choice - Equal choice) > Rank-reversal (Unequal choice - Equal choice)
z = 32
x = -6
MFG
TPJ
ACC
voxel level p < 0.005 uncorrected

Figure S9. Interaction between condition and choice. A flexible factorial analysis of the interaction between condition and choice suggests that activity in MFG, IFG, Insula, ACC, MCC, TPJ, and IPL is enhanced when people choosing the more unequal offer than choosing the more equal offer in the No Rank-reversal condition, but not in the Rank-reversal condition. Significant clusters were thresholded with small volume correction voxel-level *p*(FWE) < 0.05. For visualization, clusters were thresholded at voxel-level *p* < 0.005 uncorrected.

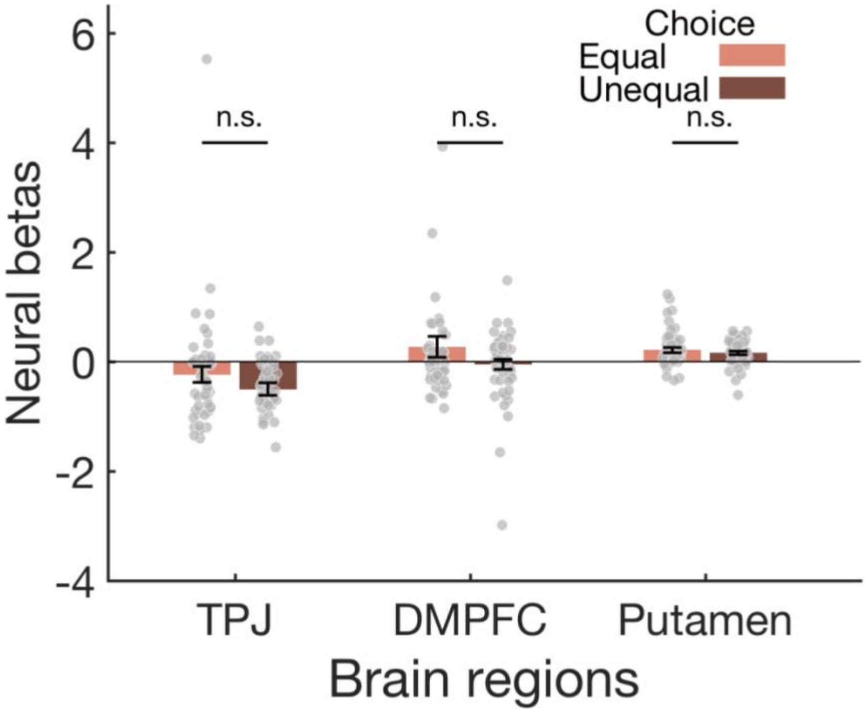


Figure S10. No group-level difference in neural responses between choosing the unequal offer and choosing the equal offer in the Rank-reversal condition. For regions whose activity were shown to be associated with model parameters (i.e., α and β) in the contrast of "Rank-reversal: unequal choice > equal choice", there was no significant difference in their responses between choosing the unequal offer and choosing the equal offer in the Rank-reversal condition at group level. Each dot represents one participant, and error bars represent the SEMs. n.s., not significant.