OXFORD

Social Cognitive and Affective Neuroscience, 2018, 578–589

doi: 10.1093/scan/nsy040
Advance Access Publication Date: 7 June 2018
Original article

Spreading inequality: neural computations underlying paying-it-forward reciprocity

Yang Hu,[1] Lisheng He,[2] Lei Zhang,[3] Thorben Wölk,[1] Jean-Claude Dreher,[4] and Bernd Weber[1,5]

[1]Center for Economics and Neuroscience, University of Bonn, 53127 Bonn, Germany, [2]Warwick Business School, The University of Warwick, Coventry CV4 7AL, UK, [3]Institute for System Neuroscience, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany, [4]Neuroeconomics, Reward and Decision Making Laboratory, Institut des Sciences Cognitives Marc Jeannerod, CNRS, 69675 Bron, France, and [5]Department of Epileptology, University Hospital Bonn, 53127 Bonn, Germany

Correspondence should be addressed to Yang Hu, Neuroeconomics, Reward and Decision Making Laboratory, Institut des Sciences Cognitives Marc Jeannerod, CNRS, 69675 Bron, France. E-mail: huyang200606@gmail.com

Present address: Lisheng He, Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA

Lisheng He and Lei Zhang contributed equally to this study.

Abstract

People tend to pay the generosity they receive from a person forward to someone else even if they have no chance to reciprocate directly. This phenomenon, known as paying-it-forward (PIF) reciprocity, crucially contributes to the maintenance of a cooperative human society by passing kindness among strangers and has been widely studied in evolutionary biology. To further examine its neural implementation and underlying computations, we used functional magnetic resonance imaging together with computational modeling. In a modified PIF paradigm, participants first received a monetary split (i.e. greedy, equal or generous) from either a human partner or a computer. They then chose between two options involving additional amounts of money to be allocated between themselves and an uninvolved person. Behaviorally, people forward the previously received greed/generosity towards a third person. The social impact of previous treatments is integrated into computational signals in the ventromedial prefrontal cortex and the right temporoparietal junction during subsequent decision making. Our findings provide insights to understand the proximal origin of PIF reciprocity.

Key words: paying-it-forward (PIF) reciprocity; inequality; model-based fMRI; ventromedial prefrontal cortex (vmPFC); right temporoparietal junction (TPJ)

Introduction

Pay-it-forward (PIF) reciprocity (also called generalized reciprocity or 'upstream' indirect reciprocity) refers to the phenomenon that even in completely anonymous interactions, an individual, once being helped by a stranger, may transmit the helping behavior to an uninvolved third person (Nowak and Sigmund, 2005; Pfeiffer *et al.*, 2005). Despite being observed in experimental settings of both humans (Bartlett and DeSteno, 2006; Gray *et al.*, 2014; Watanabe *et al.*, 2014) and non-human species (Rutte and Taborsky, 2007), PIF reciprocity is found much less likely to evolve in a well-mixed population but only in small groups (Rankin and Taborsky, 2009) and not easy to interpret from an evolutionary perspective (Pfeiffer *et al.*, 2005; Rand and Nowak, 2013). In spite of its unclear evolutionary

Received: 20 October 2017; **Revised:** 15 May 2018; **Accepted:** 29 May 2018

mechanism, PIF reciprocity plays an important role, especially for human beings, in passing kindness and thus constructing a social norm system in modern societies, which are more and more characterized by greater mobility and substantial interactions among strangers.

Previous behavioral studies on PIF reciprocity mainly focused on helping behavior. Bartlett and DeSteno (2006) found that besides returning actions back to the benefactor, people even devoted more time to helping a stranger finish a tedious problem-solving task (Bartlett and DeSteno, 2006). Another behavioral study showed that people who are being helped would transfer more money not only to the benefactor but also to a stranger (DeSteno *et al.*, 2010). Recent research on PIF reciprocity extends its scope from helping behavior to fairness-related behaviors (Gray *et al.*, 2014). For instance, in a series of experiments, Gray *et al.* (2014) revealed that participants receiving a greedy treatment (i.e. either monetary splits or workload distribution) behaved more selfishly to a third person, while they showed more kindness after being treated equally or generously in previous interactions. Similarly, people receiving an unfair (*vs* fair) offer were found to share less money with an unrelated person (Wu *et al.*, 2015). These findings indicate that people not only forward kind behaviors but also transmit less socially favored behaviors such as unfairness (see also Strang *et al.*, 2016). However, neurobiologically, it is still unclear how people integrate fairness-related information from previous interactions into subsequent computations and decisions.

To address the above questions, we applied functional magnetic resonance imaging (fMRI) in combination with computational modeling using a modified PIF design. In particular, participants first received a monetary split either from a real human player or a computer. They then had to choose one out of two split options of an additional monetary amount between themselves and a third person. One option was always a fixed equal payoff, whereas the other one was unequal, favoring either the participant or the third person. The rationale behind this setup was based on previous evidence suggesting that people's other-regarding preference depends on the payoff allocation between themselves and their matched partners (i.e. advantageous or disadvantageous inequality) (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Tricomi *et al.*, 2010). To further capture and quantify the individual's intrinsic motivation for behaving altruistically in both inequality domains, we adopted the Fehr–Schmidt model (Fehr and Schmidt, 1999), which has been widely applied to a various tasks measuring fairness-related preference (Morishima *et al.*, 2012; Sáez *et al.*, 2015; Gao *et al.*, 2018).

Based on previous findings by Gray *et al.* (2014), we expected that after being treated greedily people would be more likely to behave selfishly to the next person. In contrast, after being treated generously, they should be more likely to forward the generosity by sacrificing their own interest. Moreover, recent studies suggest that the emotion of gratitude generates the PIF reciprocity and serves as its psychological basis (Chang *et al.*, 2012). Given the original paper by McCullough *et al.* (2001), one of the prerequisites for people to feel gratitude is that the actor's generosity has to be intentional rather than accidental (McCullough *et al.*, 2001). Based on these literature, together with other studies showing the social-specific effect in social decision-making tasks (i.e. dictator game or ultimatum game with a human partner or a computer) (Sanfey *et al.*, 2003; Ruff *et al.*, 2013), we hypothesize that such inequality-dependent effect on behaviors should only appear (or be more pronounced) when participants receive the money from a human partner rather than a computer.

More importantly, the present study aimed to explore how the brain computes signals integrating previous social interactions into subsequent decisions. To this end, we took a model-based fMRI approach (O'Doherty *et al.*, 2007) building on the assumption that when confronted with a choice, the brain computes a subjective value of each option and then a decision is made by comparing these values (Padoa-Schioppa, 2011; Levy and Glimcher, 2012). Based on two lines of research, our model-based analyses mainly focused on the following two brain regions. The first region is the ventromedial prefrontal cortex (vmPFC), which has been shown as the key hub for subjective value computation during decision-making process in either neuroimaging studies (Bartra *et al.*, 2013; Clithero and Rangel, 2013) or human lesion studies (Koenigs and Tranel, 2007; Krajbich *et al.*, 2009). Plus, in a recent study, Chung *et al.* first observed that vmPFC integrated the other's influence (i.e. other-conferred utility) during a risky decision task (Chung *et al.*, 2015). The other target region is the right temporoparietal junctions (TPJ), which has been generally associated with social functions such as mentalization (Frith and Frith, 2006; Schurz *et al.*, 2014; Schaafsma *et al.*, 2015). Moreover, recent literature on structural and functional brain imaging revealed the close link between the right TPJ and prosocial decision-making (Morishima *et al.*, 2012; Tusche *et al.*, 2016; Park *et al.*, 2017), while a recent model-based fMRI study has provided further evidence for its crucial role in the computation of altruistic behaviors in the context of distributive fairness (Hutcherson *et al.*, 2015). Taken together, we expected to observe the recruitment of the vmPFC as well as the right TPJ in value computation in social contexts (i.e. receiving the monetary split from a person rather than from a computer). Previous studies have suggested that the vmPFC may encode both absolute chosen value (i.e. subjective utility of chosen options) (Zhong *et al.*, 2016) and relative chosen value (i.e. utility difference between chosen and non-chosen options) (Crockett *et al.*, 2017) in the decision process of the social context. Given the mixed evidence of vmPFC and relatively scarce evidence revealing to role of the right TPJ in computing choice value in social decision-making task (Hutcherson *et al.*, 2015; Hill *et al.*, 2017), it is difficult for us to make precise predictions linking the exact region with the computation of exact type of values/utilities. Thus we explored the neural computations related to both types of values in the current study. As an extra, this would also add to the evidence regarding the brain regions that are involved in computing the two types of choice values.

In addition, the present study aimed to extend previous findings on neural correlates of inequality perception. Using the ultimatum game (Sanfey *et al.*, 2003) or other relevant paradigms (Haruno and Frith, 2010; Tricomi *et al.*, 2010; Yu *et al.*, 2014), most previous studies only used disadvantageous splits (i.e. participants received less money than the partner) as stimuli. Surprisingly, however, few studies (Roalf, 2010; Civai *et al.*, 2012) adopted advantageous splits (i.e. participants received more money than the partner). Besides, no study, to our knowledge, investigated the social-specific effect of inequality on neural processing by introducing a non-social control (i.e. computer). Based on previous work on the neural signature of unfairness (Feng *et al.*, 2015), we assumed that both types of inequality (i.e. greedy and generous split; *vs* equal split) recruit aversive (e.g. anterior insular cortices, AI) and control-related networks (e.g. lateral prefrontal cortex, lPFC; dorsal anterior cingulate cortex, dACC). We also expected to observe the involvement of the reward circuitry (e.g. ventral striatum, VS) (Haber and Knutson, 2010; Bhanji and Delgado, 2014) in response to the

generous splits bringing participants with more monetary profits. Furthermore, we hypothesized that the effect of inequality would be reflected in the right TPJ especially when people received money from a human partner (*vs* computer).

Materials and methods

Participants

Fifty healthy participants (31 females; mean age ± s.d. = 25.2 ± 3.8 years, ranging from 18 to 33 years; 4 left-handed) were recruited via online flyers for the fMRI experiment. All participants had a normal or corrected-to-normal vision and reported no prior history of psychiatric or neurological disorders. The study was approved by the ethics committee of the University of Bonn. Written informed consent was received from all participants according to the Declaration of Helsinki (*BMJ* 1991; 302: 1194). All experimental protocols and procedures were conducted in accordance with the IRB guidelines for experimental testing and were in compliance with the latest revision of the Declaration of Helsinki. In addition, two independent groups of participants were recruited for the online task (i.e. online Groups A and B, each with 50 online participants playing the role of Players A and B, respectively) in which we collected real decisions that were used for the later fMRI experiment as stimuli (see Supplementary data for details).

fMRI task

We adopted and modified the PIF paradigm based on the study of Gray *et al.* (2014) in the current fMRI study. Notably, there were two key modifications. First, we included a non-social control such that participants received a monetary split either from a real human player (i.e. Player A) or a computer, which helps us to dissociate the social-specific components of PIF reciprocity. Another key modification was that we replaced the free transfer mode in the original version with a binary choice scheme. By using the free transfer mode as in previous studies, participants would be endowed with a certain amount of money and then asked to make a transfer to a third person whose initial payoff is 0. As a consequence, the payoff context is always advantageous to participants (vs the third person), which makes it impossible to examine people's decisions and other-regarding preference in a disadvantageous inequality context.

Hence, an event-related fMRI design was adopted with two within-subject factors, namely 'partner' (i.e. human/computer) and 'split' (i.e. greedy/equal/generous). Each 'partner × split' condition consisted of 24 trials (i.e. 144 trials in total). Specifically, each trial consisted of two independent dictator games (see Figure 1). In Game 1, participants played the role of the recipient, who received a certain monetary split with a fixed total amount of €10 from a proposer, who could be either a real person (i.e. a Player A in the online Group A; indicated by the initials; varied across trials) or a computer. Crucially, the proposer could offer a greedy (i.e. giving less than €5), equal (i.e. giving €5) or generous (i.e. giving more than €5) split to the participant. This period lasted 2 s, which was followed by an inter-stimulus interval showing a jittered fixation cross (mean = 3 s; 1 – 5 s). In Game 2, participants played the role of the proposer. They were presented with two options of money splits between him-/herself and an uninvolved person (i.e. a Player B in the online Group B; indicated by the initials; varied across trials). Particularly, one of the options was always equal with a fixed payoff earning €5 for both participants and Player B. The alternative option was unequal, which either earned more than €5 for the participant (and less than €5 for Player B; advantageous context), or earned less than €5 for the participant (and more than €5 for Player B; disadvantageous context). Participants were asked to select one of the two options within 4 s, by pressing the corresponding buttons on the button box with their left or right index fingers. Once the decision was made, a magenta frame appeared to indicate the chosen option for the remaining time (i.e. 4 s minus the decision time). If they failed to respond within 4 s or made an unrealistically fast decision (i.e. decision time <200 ms), a warning screen was presented for 1 s. As a consequence, participants would not obtain any money in these trials. Each trial ended with an inter-trial interval showing another jittered fixation cross (mean = 5 s; 3 – 7 s). To increase the variation of the stimuli and maintain participants' attention during the experiment, we added a uniformly distributed random fluctuation to payoffs (e.g. €8.08/€1.92; see Supplementary Tables S1 and S2 for the full list of stimuli in both games). The 'partner × split' conditions in Game 1 were pseudo-randomly presented to participants. Besides, a certain split in Game 1 was randomly paired with an unequal option in Game 2. Such specific pairs were kept the same for human partner and computer within each participant, ruling out the differential payoffs between human and computer condition.

At the end of the experiment, one trial was randomly selected, and the payoff in that trial would be used as their final payment (i.e. the monetary amount they received from a Player A in the online Group A or the computer in Game 1, plus the amount they kept in the selected options in Game 2). Importantly, participants were informed that their chosen decisions would also match the corresponding Player B (in the online Group B) in Game 2 and cause real monetary consequences for them. Importantly, all these procedures above were real, following the rule of no deception widely used in the behavioral economic studies.

All stimuli were presented using Presentation v14 (Neurobehavioral Systems, Inc., Albany, CA, USA) on a 32″ liquid crystal display (NordicNeuroLab, Bergen, Norway) outside the scanner with a resolution of 800 × 600 pixels, using a mirror system attached to the head coil.

Procedure

On the day of scanning, participants were first given the instructions about the experimental task and informed about the online part. Next, they completed a series of comprehension questions to ensure that they fully understood the task. Before the incentivized fMRI task, participants completed a practice session (i.e. no more than five trials) to get familiar with the paradigm as well as the button responses in the scanner. The fMRI task included one functional session lasting ~40 min, which was followed by a 6-min structural scan. In the end, participants received, via bank transfer, a €10 show-up fee, a €5 bonus for limiting their head motion during fMRI scanning (which, if exceeding 3 mm, would not be paid), and a decision-dependent payoff (range: €2–18).

Data acquisition

Participants' responses in the scanner were collected via an MRI-compatible response device (NordicNeuroLab). The imaging data were acquired on a 3-Tesla Siemens Trio MRI system (Siemens, *Erlangen, Germany*) with a 32-channel head coil at the Life & Brain

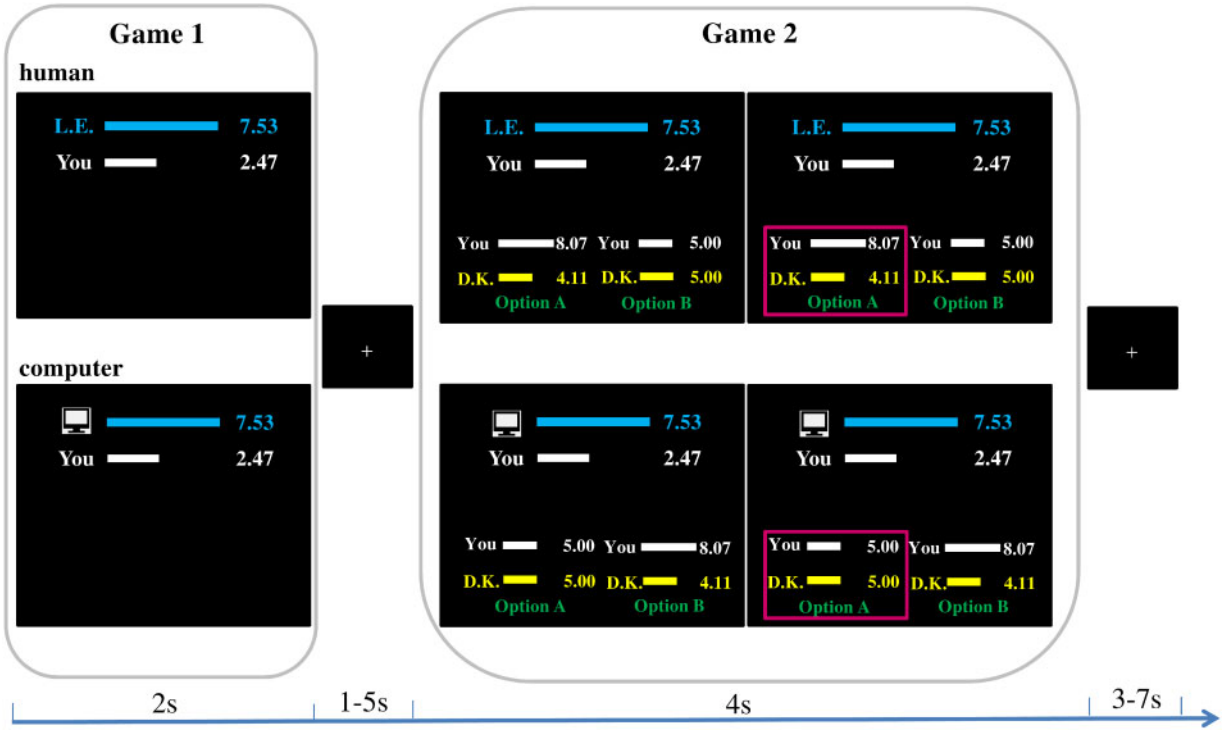

Fig. 1. Example of trial procedure in the scanner. Each trial included two independent dictator games (labeled as 'money split game'). In Game 1, participants played the role of the recipient and received a certain split of money (in total €10) either from a real Player A (indicated by the initials, e.g. L.E.) or the computer (indicated by an icon), which lasted 2 s. The split could be greedy (i.e. gaining less than €5, as shown here), equal (i.e. gaining €5) or generous (i.e. gaining more than €5) for participants. Following a jittered fixation cross (1–5 s), participants in Game 2 played the role of the proposer and decided between two options about splitting an additional amount of money with another person (i.e. Player B) within 4 s. One of the options was always equal with the fixed payoff of €5 for each side; the other option was unequal, causing either advantageous (i.e. earning more than Player B, as shown here) or disadvantageous (i.e. earning less than Player B) inequality status for the participants. Once the decision was made, a magenta frame appeared to indicate the chosen option for the remaining time. The trial ended with another jittered fixation cross (3–7 s).

Center, University Hospital Bonn. The functional scans were acquired using a T2*-weighted echo planar imaging (EPI) pulse sequence employing a BOLD contrast (TR = 2500 ms; TE = 30 ms; flip angle = 90°) in 37 axial slices (FOV = 192 × 192 mm^2, matrix = 96 × 96, thickness = 3 mm, in-plane resolution = 2 × 2 mm^2) covering the whole brain volume. Slices were axially oriented along the AC–PC plane and acquired in ascending order. A high-resolution structural T1-weighted image was also collected for every participant using a 3D MRI sequence (TR = 1660 ms; TE = 2.75 ms; flip angle = 9°; matrix = 320 × 320; slice thickness = 0.8 mm; FOV = 256 × 256 mm^2).

Data analyses

Data from two participants were excluded due to excessive head movements (> 3 mm), thus later analyses were performed based on the data of remaining 48 participants (30 females).

Behavioral analyses

Behavioral data (i.e. choice and decision time) were analyzed by mixed-effect regressions using R (http://www.r-project.org/; for a summary of descriptive statistics, see Table 1; also see Supplementary data for details). Participants' decisions in Game 2 were labeled as generous if they chose the option relatively earning less (more) for themselves (Player B), and otherwise selfish. To quantify each participant's degree of aversion to either advantageous or disadvantageous inequality (in Game 2), we adopted the Fehr–Schmidt model (Fehr and Schmidt, 1999) as the base model (i.e. m1) and established alternative models by taking into account different factors in Game 1 (i.e. *partner*, *split* or both factors; m2–m5). Model comparison and parameter estimates were performed with a hierarchical Bayesian approach via the 'hBayesDM' package (Ahn *et al.*, 2017; see Supplementary data for details).

fMRI analyses

Functional imaging data were analyzed using SPM 8 (Wellcome Trust Centre for Neuroimaging, University College London, London, UK). For each participant, the preprocessing of the functional data followed the common pipeline: (i) the first three volumes were discarded to allow for the stabilization of the BOLD signal; (ii) EPI images were realigned to the first volume to correct motion artifacts and then corrected for slice timing; (iii) the structural T_1 image was co-registered to the mean EPI images and then segmented into white-matter, grey-matter and cerebrospinal fluid to generate normalization parameters to MNI space; (iv) all EPI images were normalized to the MNI space, resampled with a 2 × 2 × 2 mm^3 resolution, based on parameters generated in the previous step, and then smoothed using an 8-mm isotropic full width half maximum Gaussian kernel; (v) high-pass temporal filtering was performed with a cut-off value of 128 s to remove low-frequency drifts.

For each participant, we established two GLMs to investigate the effect of 'partner' and 'split' (Game 1) on computation-relevant neural signals during decision-making period in Game 2. Specifically, GLM1 contained 12 regressors of interest: the

Table 1. Summary of descriptive statistics in Game 2

		Generous		Selfish	
		Human	Computer	Human	Computer
Choice proportion (%; mean±s.d.)[a]	Greedy	18.7±18.9	20.1±19.1	81.2±17.9	79.6±19.2
	Equal	23.4±20.8	23.7±21.8	76.2±20.5	76.2±21.7
	Generous	25.7±24.5	25.4±23.8	74.1±24.4	73.9±23.3
Decision time	Greedy	1756.5±505.6	1701.0±442.1	1426.8±358.4	1400.3±340.6
(ms; mean±s.d.)	(*N*)	(39)	(36)	(48)	(48)
	Equal	1647.1±439.2	1710.0±484.0	1381.0±327.6	1381.8±368.4
	(*N*)	(37)	(34)	(48)	(48)
	Generous	1678.4±494.7	1630.7±455.8	1451.4±437.9	1433.9±414.0
	(*N*)	(38)	(38)	(47)	(48)

Note: we first calculated the individual-level mean (±s.d.) choice proportion and decision time in terms of specific decisions for each condition, then we calculated the group-level mean (±s.d.) based on the individual mean; due to individual difference in decisions, the sample size (i.e. *N*) for each condition of specific decisions is different.

[a]The sample size for calculating the choice proportion is always 48.

onsets of the decision period in Game 2 sorted by the 'partner × split' treatment in Game 1 (i.e. human: greedy, equal, generous; computer: greedy, equal, generous; duration equals the actual decision time) as well as the corresponding parametric modulators (PMs) with absolute chosen values derived from the winning model. The *nuisance* regressor consisted of onsets of the monetary split presentation in Game 1 (duration equals 2 s), onsets of the decision period with too fast (duration equals actual decision time) or no response (duration equals 4 s), as well as the warning feedback (duration equals 1 s) in Game 2, which were considered as events of no interest. GLM2 was built in the same way as GLM1 except that we used the relative chosen values derived from the winning model as the PMs (see Supplementary data for details).

In addition, we established GLM3 to examine the effect of 'partner' and 'split' on the neural responses in Game 1. Thus, we included the following six regressors of interest, namely onsets of monetary split presentation sorted by the 'partner × split' treatment in Game 1 with the duration of 2 s. Besides, a regressor modeling events of no interest (i.e. nuisance) was also included, which contained onsets of decisions (duration equals the actual decision time; for trials of no response, duration equals 4 s) and the warning feedback due to fast response or no response (duration equals 1 s) in Game 2.

For all above GLMs, the canonical hemodynamic response function was used to model the fMRI signal. Besides, the six movement parameters were added to all models as covariates to account for motion artifact. Linear contrasts of regression coefficients of these regressors of interest (vs implicit baseline; PMs used in GLM1 and GLM2) were computed at the individual subject level in each GLM and then forwarded to group-level random-effect analyses. Particularly, a 2 × 3 within-subject flexible factorial ANOVA model was used for contrasts in three GLMs, respectively, each with the within-subject factors of the 'partner', 'split' and 'partner × split' included. Pair-wise *t*-tests were performed to unpack the simple effect of 'split' and 'partner × split' interaction in above analyses.

We adopted a whole-brain corrected threshold of $P < 0.05$ at the cluster-level controlling for family-wise error rate with an uncorrected voxel-level threshold of $P < 0.001$ as the cluster-defining threshold (Eklund *et al.*, 2016) for all results above. Additionally, a small volume correction (SVC) was conducted within the pre-defined coordinate-based mask of the right TPJ based on (Hutcherson *et al.*, 2015). To illustrate the effect of PMs in GLM1 and GLM2, we adopted the 'rfxplot' toolbox (http://rfxplot.sourceforge.net/; Gläscher, 2009).

Results

Behavioral results

Choice. Compared with the 'null' model, the 'main-effect-only' model provided a significant better fit to the choice data [likelihood ratio test, LRT: $\chi^2(3) = 34.30$, $P < 0.001$], showing that participants were more likely to choose the generous option after being treated equally (odds ratio = 1.37, $b = 0.31$, $P < 0.001$) or generously (odds ratio = 1.58, $b = 0.46$, $P < 0.001$), both compared with receiving the greedy split. A trend-to-significant increase on likelihood of choosing generous options was detected when participants received the generous split (*vs* equal; odds ratio = 1.16, $b = 0.15$, $P = 0.061$). No significant difference on choosing the generous option was observed between the human and computer condition (odds ratio = 0.96, $b = -0.04$, $P = 0.551$; Figure 2A). However, the model fit was not improved by additionally including the 'partner × split' interaction in the 'main-and-interaction' model [*vs* the 'main-effect-only' model; LRT: $\chi^2(2) = 0.70$, $P = 0.705$; see Supplementary Table S3 for details; for results on decision time, see Supplementary Figure S1 and Tables S4 and S5 for details].

Computational modeling. The hierarchical Bayesian analysis and Bayesian model comparison showed that the model (i.e. m3) that distinguished advantageous/disadvantageous inequality aversion parameters in terms of 'split' has the lowest WAIC scores (Figure 2B), suggesting that it outperformed other competing models in term of out-of-sample predictive accuracy. To further check the effect of 'split' on the degree of inequality aversion in the disadvantageous and advantageous domain, we performed mixed-effect linear regressions on the posterior mean of individual α (i.e. the parameter capturing the aversion degree to disadvantageous inequality, a.k.a., the 'envy' parameter) and β (i.e. the parameter capturing the aversion degree to advantageous inequality, a.k.a., the 'guilt' parameter) estimates from the winning model (i.e. m3). In each regression, 'split' was adopted as the fixed-effect predictor (coded as dummy

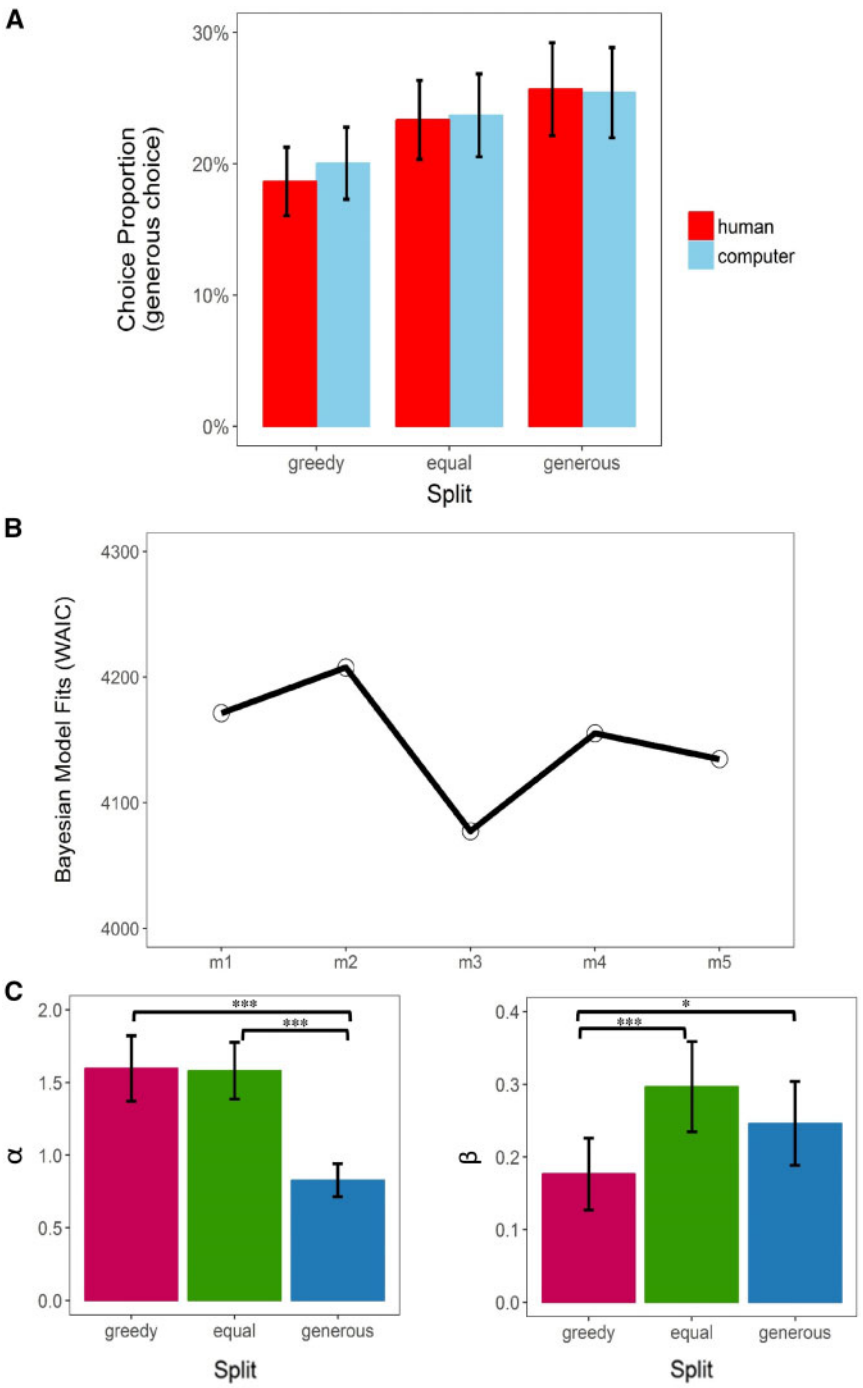

Fig. 2. Behavioral results. (A) The proportion (%) of generous choices participants made in Game 2 given different 'partner × split' conditions in Game 1. (B) Bayesian model comparisons of all five candidate models (m1–m5). The lower WAIC score indicates better out-of-sample prediction accuracy of the candidate model. Here model 3 (m3) outperforms other candidate models. WAIC, widely applicable information criterion; m1–5 = model 1–5. (C) Bar plot of the group-level mean of the posterior distribution of the estimated parameters based on the winning model (i.e. m3). α measures the degree of aversion to disadvantageous inequality (i.e. how the participant dislikes that he/she earned less than Player B); β measures the degree of aversion to advantageous inequality (i.e. how the participant dislikes that he/she earned more than Player B). Error bars refer to SEM. Significance level: ***$P < 0.001$, *$P < 0.05$.

variables; reference level: greedy or equal). Participants felt less aversive to the disadvantageous unequal option (as denoted by α) while being treated generously (*vs* greedy: mean ± s.d.: 0.83 ± 0.79 *vs* 1.60 ± 1.56, $b = -0.77$, $P < 0.001$; *vs* equal: 1.58 ± 1.35, $b = -0.75$, $P < 0.001$). On the other hand, they felt less aversive to the advantageous unequal option (as denoted by β) while being treated greedily (*vs* equal: mean ± s.d.: 0.18 ± 0.34 *vs* 0.30 ± 0.43, $b = -0.12$, $P < 0.001$; *vs* generous: 0.25 ± 0.43, $b = -0.07$, $P = 0.044$; Figure 2C; for results of non-parametric analyses, see Supplementary Figure S2).

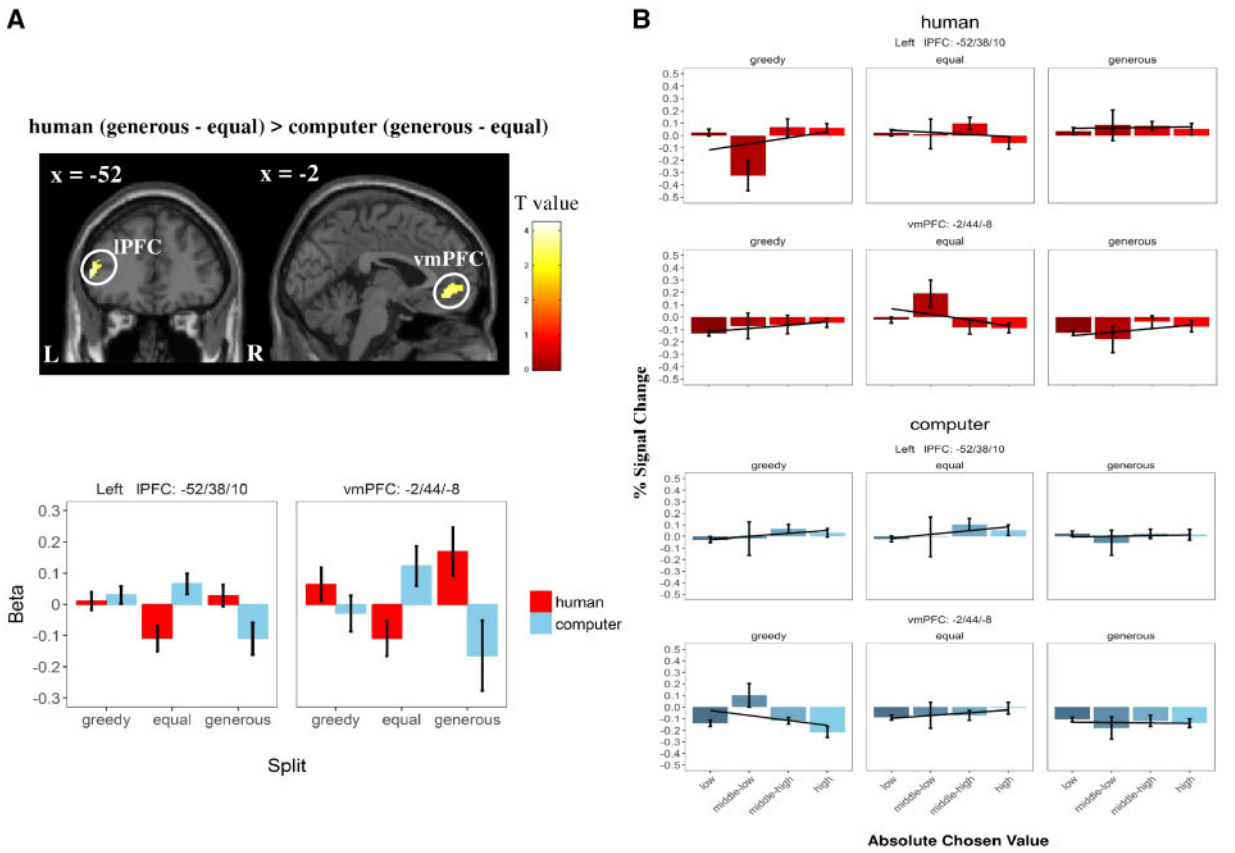

Fig. 3. Impact of previous treatment (Game 1) on neural computation of absolute chosen value (i.e. trial-wise subjective utility of chosen option; GLM 1) during decision-making process in Game 2. (A) Regions reflecting absolute chosen value while receiving generous (*vs* equal) splits from a human partner (*vs* a computer). lPFC, lateral prefrontal cortex; vmPFC, ventral medial prefrontal cortex. For display reason, β values of both regions (local peak voxels) in different conditions were extracted. Display threshold: cluster-level P (FWE-corrected) <0.05 together with voxel-level P (uncorrected) <0.001. (B) Differential modulation of absolute chosen value on left lPFC and vmPFC (local peak voxels) in different conditions. Relative chosen value is split into four bins, i.e. low (0–25%), medium–low (25–50%), medium–high (50–75%) and high (75–100%). Error bars refer to SEM; line refers to the linear fit.

Imaging results

Neuro-computations during decision-making in Game 2 (GLM1 and GLM2). In GLM1, we found a stronger 'partner×split' interaction in signals of absolute chosen values in the vmPFC and left lPFC [i.e. human: (generous–equal) > computer: (generous–equal)]. These regions showed a more positive modulation when receiving a generous (*vs* equal) split from a human partner than from the computer in Game 1 (Figure 3). Moreover, we showed in GLM2 that signals of relative chosen values in the right TPJ, extending to the posterior superior temporal sulci, were stronger while receiving the split from a human partner than from a computer (Figure 4; see Supplementary Table S6 for details). No region was detected in other main effects and interaction contrasts under the same threshold in either GLM.

Neural correlates of inequality perception in Game 1 (GLM3). Either receiving the greedy or the generous split, compared with the equal split, yielded a stronger activation in the left AI, dorsal ACC and bilateral lPFC, which was further confirmed by a conjunction analysis (i.e. greedy > equal AND generous > equal, Figure 5A; see Supplementary Table S7 for details). The comparison between receiving a generous and a greedy monetary split, on the other hand, revealed higher activation in reward-related areas, including the vmPFC and bilateral VS (i.e. generous > greedy, Figure 5B; see Supplementary Table S8 for details).

Moreover, we found an increased activity in the bilateral lPFC as well as right TPJ [peak MNI coordinates: 50, −60, 32; $t(235) = 3.76$, P(SVC-FWE) = 0.051] while participants received a monetary split from a human partner than a computer (i.e. human > computer). The reverse contrast (i.e. computer > human) only yielded activations in occipital areas. Interestingly, we observed a 'partner × split' interaction effect in the right TPJ extending to the inferior parietal lobule [IPL; peak MNI coordinates: 56, −48, 34; $t(235) = 4.06$, P(SVC-FWE) = 0.020], which responded stronger for the generous (*vs* equal) split offered by a human partner than a computer (i.e. human: [generous–equal] > computer: [generous–equal]; see Supplementary Figure S3 and Table S8 for details). No region was detected in other interaction contrasts under the same threshold.

Discussion

The current fMRI study adopted a modified PIF paradigm to investigate whether and how people spread inequality to uninvolved strangers. As predicted, receiving a greedy monetary split makes people become more selfish themselves. However, people only showed a marginally significant increase in the generosity level when they receive a generous monetary split. More intriguingly, our results from the behavioral modeling further confirm and extend the above findings by dissociating people's altruistic motivation from different inequality contexts.

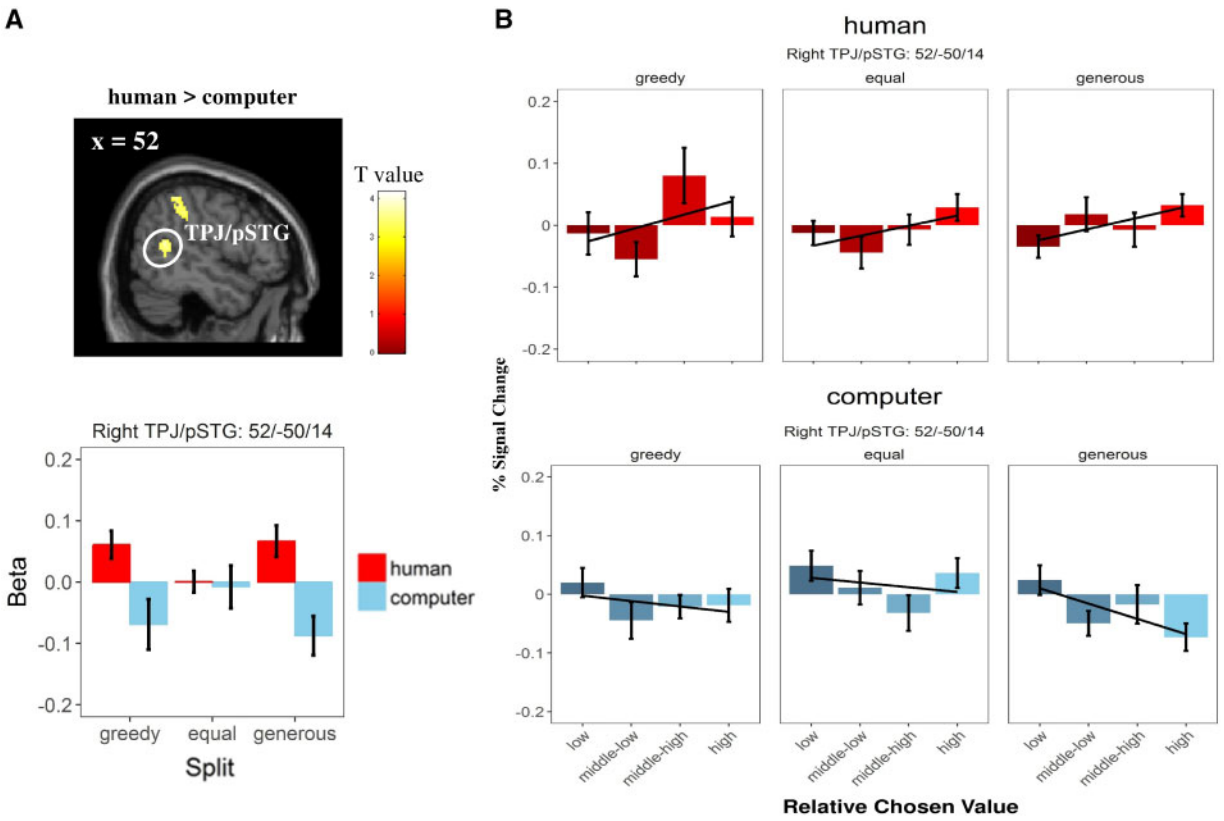

Fig. 4. Impact of previous treatment (Game 1) on neural computation of relative chosen value (i.e. trial-wise subjective utility difference between chosen and non-chosen option; GLM 2) during decision-making process in Game 2. (A) Regions reflecting relative chosen value while receiving monetary splits from a human partner (*vs* a computer). TPJ/pSTG = temporo-parietal junction/pSTG. For display reason, β values of both regions (local peak voxels) in different conditions were extracted. Display threshold: cluster-level P (FWE-corrected) < 0.05 together with voxel-level P (uncorrected) < 0.001. (B) Differential modulation of relative chosen value on right TPJ/pSTG (local peak voxels) in different conditions. Relative chosen value is split into four bins, i.e. low (0–25%), medium–low (25–50%), medium–high (50–75%) and high (75–100%). Error bars refer to SEM; line refers to the linear fit.

In the advantageous inequality context (measured by β), people became less aversive to the advantageous unequal option (i.e. the selfish option) after receiving the greedy split. In the disadvantageous inequality context (measured by α), people were less aversive to the disadvantageous unequal option (i.e. the generous option) after receiving a generous split. However, such change in disadvantageous inequality aversion did not lead to an overall increase in choosing the generous option. Presumably, this might be due to the dominance of the advantageous inequality context adopted in the design (i.e. more trials with advantageous unequal options in Game 2; see Supplementary Table S2). Taken together, our findings are in line with the asymmetry shown previously in the PIF task that participants forward the greed, rather than the generosity, the most (Gray *et al.*, 2014).

Although no behavioral difference between the human partner and the computer condition was observed, we did observe significant differences at the brain level especially during the subsequent decision-making process (Game 2). As predicted, we observed that the vmPFC integrated both social and inequality information with signals representing absolute chosen values. As a crucial region repeatedly involved in value-based decision making (Rangel and Hare, 2010; Bartra *et al.*, 2013; Clithero and Rangel, 2013) and social preference (Seo and Lee, 2012), the vmPFC is proposed to reflect the computation at the time of decision-making, i.e. reflecting subjective decision-values (Hare *et al.*, 2008; Ruff and Fehr, 2014). Human lesion studies provide causal evidence indicating the crucial role of vmPFC in value computation during economic decision making in social contexts (Koenigs and Tranel, 2007; Krajbich *et al.*, 2009). Not only consistent with previous findings, our results furthermore showed that computational signals of the vmPFC are found to be stronger while participants receive a generous split from a human partner rather than a computer, indicating its role in integrating social-related contextual information. This result is not only in line with the general findings in value-based decision making (Padoa-Schioppa, 2011; Levy and Glimcher, 2012) but also fits a previous study linking the vmPFC to computational signals during charitable decision-making processes (Hare *et al.*, 2010). Apart from vmPFC, we also found that the same contrast yielded stronger activation in the left lPFC. Regarded as the key hub in executive control (Miller and Cohen, 2001), the lPFC also plays a critical role in encoding decision values (Rangel and Hare, 2010), including social contexts (Sanfey, 2007; Rilling and Sanfey, 2011; Ruff and Fehr, 2014; Strang *et al.*, 2014). For instance, Crockett *et al.* (2017) have shown that, while people evaluated the allocation between a financial gain for themselves and somatosensory pain for either themselves or others, the left lPFC encoded the benefit gained from harming others instead of the self (Crockett *et al.*, 2017). Our results provide further evidence for the role of the dlPFC in computing other-regarding decision values.

We also revealed a social-specific effect on computations in the right TPJ during the decision-making period. In particular,

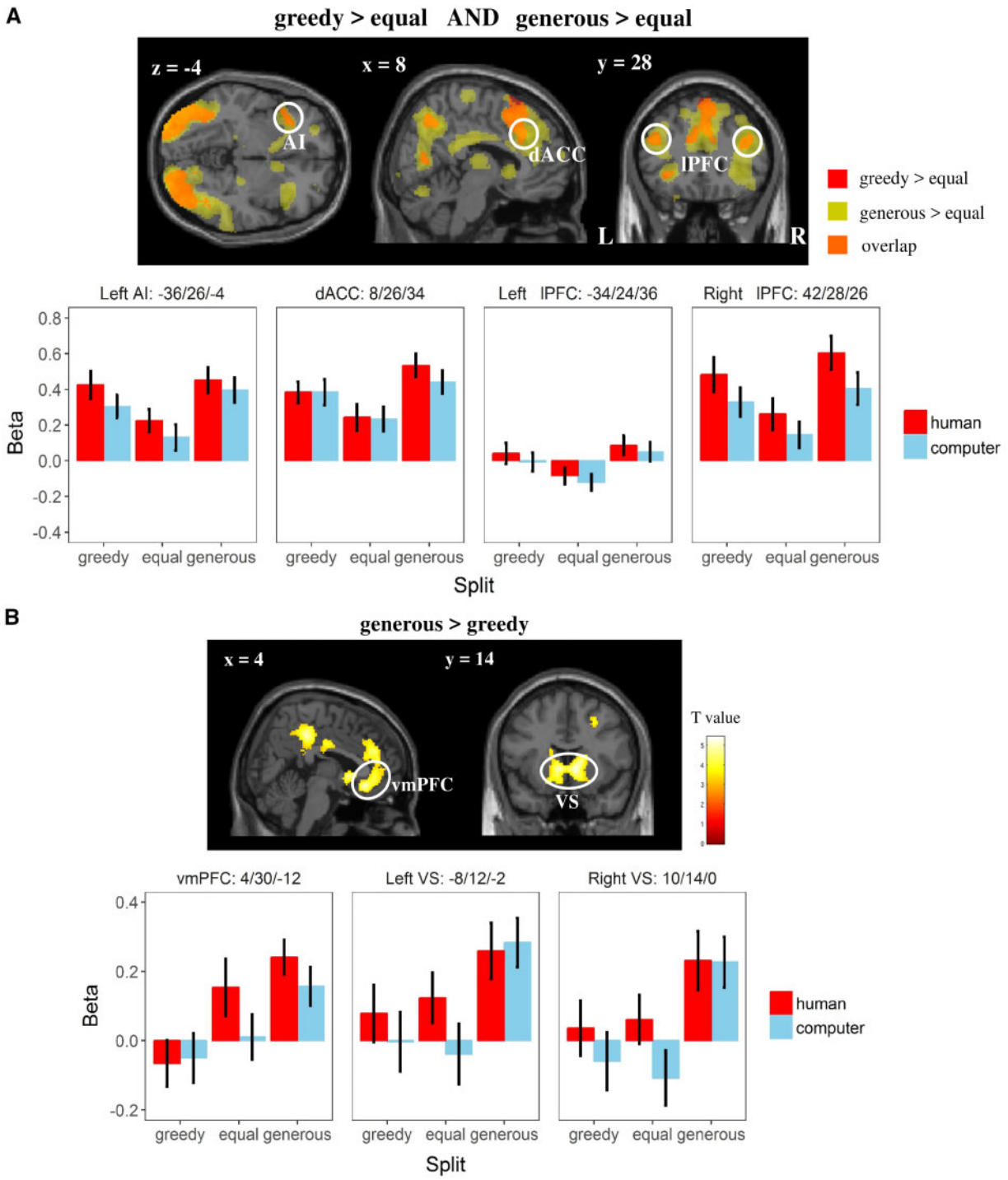

Fig. 5. The effect of *split* on neural correlates in Game 1 (GLM 3). (A) Shared neural representation of both types of inequality. AI, anterior insula; dACC, dorsal anterior cingulate cortex. (B) Regions showing stronger activation to generous (*vs* greedy) split. vmPFC, ventral medial prefrontal cortex; VS, ventral striatum. For display reason, β values from the local peak voxel were extracted in all analyses above. Error bars refer to SEM. display threshold: voxel-level P (uncorrected) <0.001, $k = 150$.

decision-relevant computational signals of relative chosen values (i.e. subjective utility differences between the chosen and non-chosen option) in the right TPJ, extending to posterior superior temporal gyrus (pSTG), are stronger if the previous partner is a human (vs computer). Consistently, the right TPJ (especially, the posterior part) is also found to respond stronger to the human partner than to the computer when people receive the monetary split (especially, the generous split; in Game 1). A large amount of literature has closely associated the right TPJ with social cognition (Decety and Lamm, 2007; Van Overwalle, 2009), especially theory-of-mind/mentalizing (Schurz *et al.*, 2014; Schaafsma *et al.*, 2015; Tusche *et al.*, 2016). Regarding the decision-making process, a previous study revealed a distinct role of the TPJ in predicting decisions when people interact with a human partner rather than a computer in an incentivized-strategic poker game (Carter *et al.*, 2012).

Furthermore, recent studies in the field of decision neuroscience unveil the role of the right TPJ in other-regarding preference (e.g. generosity), providing both anatomical and functional evidence (Morishima *et al.*, 2012; Strombach *et al.*, 2015; Park *et al.*, 2017). Notably, none of the above studies adopts the computer as a non-control condition. Thus, our results further confirm the role of the right TPJ in capturing other-regarding preferences (e.g. altruism and generosity) within the social context. Taken together, our neuroimaging findings suggest that people implicitly take into account the social component of the previous treatment affecting later computations during the decision-making process, although this effect might not be sufficiently intense to boost a behavioral difference.

The present study additionally examines the common and differential neural representation of different forms of inequality (in Game 1). In line with our predictions, receiving either of the two unequal compared with the equal split, activated the anterior insula (AI; especially, the left part), the dorsal ACC as well as the bilateral lPFC regardless of the partner. These results not only replicate previous findings of neural responses toward disadvantageous inequality (i.e. receiving greedy monetary splits) (Sanfey *et al.*, 2003) but also extends our knowledge of inequality perception to the advantageous domain. As indicated in a recent meta-analysis on neural correlates of fairness-related decision making based on the ultimatum game (Feng *et al.*, 2014), stronger activation in AI toward (disadvantageous) unequal offers might tackle a cognitive heuristic to detect norm violations, including either type of inequality. Such a norm violation causes a potentially motivational conflict between self-interest and the social norm and also an emotional conflict, which is presumably monitored or regulated by the dACC (Botvinick *et al.*, 2004) and resolved by the lPFC (Knoch *et al.*, 2006; Guo *et al.*, 2014). Interestingly, previous studies examining both types of inequality did not report the above regions (Haruno and Frith, 2010; Tricomi *et al.*, 2010; Yu *et al.*, 2014). We argue that such differences in neural activation might be due to the task difference. Specifically, in these cases, the unequal monetary split presented to participants was chosen by the experimenter, unlike in our case or the ultimatum game where it was decided by the partner. This might lead to less emotional conflict and lower recruitment of regions like the dACC or lPFC during this process. In addition, the direct comparison between the neural activation during the period of receiving the generous and greedy splits reveals the involvement of reward-related areas, including the vmPFC and VS, again as predicted. This finding can be explained either by the effect of monetary outcomes, by the intention of kindness, such as a 'warm glow' (Andreoni, 1990) or gratitude (Nowak and Roch, 2007), or a mixture of these two factors. However, it is difficult to decide among those alternatives because no 'partner × split' interaction is observed in the above regions.

Several caveats are worth notifying regarding the interpretation of results as well as the task design of the current study. To begin with, we acknowledge that the self-focusing motivation, rather than indirect reciprocity, serves as an alternative explanation which predicts the behavioral results equally well (i.e. I am selfish/generous to the other just because I have less/more money, rather than you treat me greedy/generous). The similar behavioral pattern observed in response to human partner and computer might be due to the fact that a real interpersonal interaction (e.g. introducing confederates to participants before the task) was not involved in the present paradigm, as was the case in another study investigating social decision making (Zhang *et al.*, 2016). Alternatively, the similar PIF behavioral pattern observed in the computer condition might challenge the social-specific feature of PIF reciprocity. Can PIF reciprocity happen in non-social context? We do not have a clear answer but, generally speaking, this idea is supported by previous studies which provide the empirical link that winning/earning money induces positive emotion/mood (Dunn *et al.*, 2011), and the latter makes people inclined to behave generously (Lyubomirsky *et al.*, 2005). More interestingly, a recent behavioral study showed that participants paid the reciprocity back or forward dependent not only on the generosity degree of givers but also on their wealth level (i.e. participants received more money in general from wealthier givers), indicating that the monetary reward per se indeed exerts an impact (Hackel and Zaki, 2018). In addition, another possible explanation coming from the view of social influence (Chung *et al.*, 2015; Leong and Zaki, 2018) also fits the current results, namely that participants forward the generosity/greed to the third person merely because they observe a similar behavior in others. This viewpoint enlightens future studies which aim at directly comparing behaviors (and neural activities) between these two conditions (i.e. receiving Player A's split *vs* observing Player A's behavior to others). Besides, a baseline condition prior to the fMRI design (i.e. participants only make choices in Game 2 without any treatment in Game 1) should be considered, which enables us to obtain the baseline of people's social preference, and thus to calculate the accurate behavioral PIF effect. All in all, future studies are needed to address these limitations and issues.

In real life, people are always inclined to pass on the kindness to someone else if someone treats them nicely. Here, we extend the situation to inequality and provide the first empirical evidence, to our knowledge, on the neuro-computational mechanisms underlying such PIF reciprocity by taking a model-based fMRI approach (O'Doherty *et al.*, 2007), which becomes an emergent trend in the field of social neuroscience due to its contribution in providing the mechanistic account for social decision-making (Dunne and O'Doherty, 2013). Our findings indicate that brain regions involved in value-representation and social cognition integrate the social-specific (un)equal treatment from a stranger and drives the subsequent other-regarding decisions to another uninvolved person. Looking from a different angle, our findings also inspire future studies on people with mental disorders. For instance, we might expect that people with autistic spectrum disorder would blur the distinction between a human partner and a computer both at the behavioral level and the neural level due to the impairment in their mentallizing ability (Baron-Cohen *et al.*, 1985). Taken together, we believe that these results not only improve our understanding of the neural correlates of PIF reciprocity in the context of (in)equality, but also have important implications for more broad areas such as education and social policy.

Supplementary data

Supplementary data are available at *SCAN* online.

Acknowledgements

We would like to thank Sima Hakimi for proofreading of the early draft.

Funding

Y.H. was supported by the State Scholarship Fund of the China Scholarship Council (CSC; No. 201306140034). L.Z. was

partially supported by the German Research Foundation (DFG GRK 1247), the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research (Grant01GQ1006) and the Research Promotion Fund (FFM) for young scientists of the University Medical Center Hamburg-Eppendorf. J-C.D. was funded by the grant 'ANR-NSF CRCNS "SOCIAL_POMDP" n°16-NEUC'. This work was performed within the framework of the LABEX ANR-11-LABEX-0042 of Université de Lyon, within the program 'Investissements d'Avenir' (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

Conflict of interest. None declared.

References

Ahn, W.-Y., Haines, N., Zhang, L. (2017). Revealing neuro-computational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry*, **1**, 24–57.

Andreoni, J. (1990). Impure altruism and donations to public goods: a theory of warm-glow giving. *The Economic Journal*, **100**(401), 464–77.

Baron-Cohen, S., Leslie, A.M., Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, **21**(1), 37–46.

Bartlett, M.Y., DeSteno, D. (2006). Gratitude and prosocial behavior: helping when it costs you. *Psychological Science*, **17**(4), 319–25.

Bartra, O., McGuire, J.T., Kable, J.W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, **76**, 412–27.

Bhanji, J.P., Delgado, M.R. (2014). The social brain and reward: social information processing in the human striatum. *Wiley Interdisciplinary Reviews: Cognitive Science*, **5**(1), 61–73.

Bolton, G.E., Ockenfels, A. (2000). ERC: a theory of equity, reciprocity, and competition. *American Economic Review*, **90**(1), 166–93.

Botvinick, M.M., Cohen, J.D., Carter, C.S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, **8**(12), 539–46.

Carter, R.M., Bowling, D.L., Reeck, C., Huettel, S.A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, **337**(6090), 109–11.

Chang, Y.-P., Lin, Y.-C., Chen, L.H. (2012). Pay it forward: gratitude in social networks. *Journal of Happiness Studies*, **13**(5), 761–81.

Chung, D., Christopoulos, G.I., King-Casas, B., Ball, S.B., Chiu, P.H. (2015). Social signals of safety and risk confer utility and have asymmetric effects on observers' choices. *Nature Neuroscience*, **18**(6), 912–6.

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

Clithero, J. A., Rangel, A. (2013). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, **9**(9), 1289–302.

Crockett, M.J., Siegel, J.Z., Kurth-Nelson, Z., Dayan, P., Dolan, R.J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, **20**(6), 879–85.

Decety, J., Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *The Neuroscientist*, **13**(6), 580–93.

DeSteno, D., Bartlett, M.Y., Baumann, J., Williams, L.A., Dickens, L. (2010). Gratitude as moral sentiment: emotion-guided cooperation in economic exchange. *Emotion*, **10**(2), 289.

Dunn, E.W., Gilbert, D.T., Wilson, T.D. (2011). If money doesn't make you happy, then you probably aren't spending it right. *Journal of Consumer Psychology*, **21**(2), 115–25.

Dunne, S., O'Doherty, J.P. (2013). Insights from the application of computational neuroimaging to social neuroscience. *Current Opinion in Neurobiology*, **23**(3), 387–92.

Eklund, A., Nichols, T. E., Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, **113**, 7900–5.

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

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

Frith, C.D., Frith, U. (2006). The neural basis of mentalizing. *Neuron*, **50**(4), 531–4.

Gao, X., Yu, H., Saez, I., *et al.* (2018) Spatial gradient in activity within the insula reflects dissociable neural mechanisms underlying context-dependent advantageous and disadvantageous inequity aversion. *bioRxiv*, https://doi.org/10.1101/243428.

Gläscher, J. (2009). Visualization of group inference data in functional neuroimaging. *Neuroinformatics*, **7**(1), 73–82.

Gray, K., Ward, A.F., Norton, M.I. (2014). Paying it forward: generalized reciprocity and the limits of generosity. *Journal of Experimental Psychology: General*, **143**(1), 247.

Guo, X., Zheng, L., Cheng, X., *et al.* (2014). Neural responses to unfairness and fairness depend on self-contribution to the income. *Social Cognitive and Affective Neuroscience*, **9**(10), 1498–505.

Haber, S.N., Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, **35**(1), 4–26.

Hackel, L. M., Zaki, J. (2018). Propagation of Economic Inequality Through Reciprocity and Reputation. *Psychological science*, **29**(4), 604–13.

Hare, T.A., Camerer, C.F., Knoepfle, D.T., O'Doherty, J.P., Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *The Journal of Neuroscience*, **30**(2), 583–90.

Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W., Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, **28**(22), 5623–30.

Haruno, M., Frith, C.D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nature Neuroscience*, **13**(2), 160–1.

Hill, C.A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J.P., Ruff, C.C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*, **20**(8), 1142.

Hutcherson, C., Bushong, B., Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, **87**(2), 451–62.

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

Koenigs, M., Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage:

evidence from the Ultimatum Game. *Journal of Neuroscience*, **27**(4), 951–6.
Krajbich, I., Adolphs, R., Tranel, D., Denburg, N.L., Camerer, C.F. (2009). Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *Journal of Neuroscience*, **29**(7), 2188–92.
Leong, Y.C., Zaki, J. (2018). Unrealistic optimism in advice taking: a computational account. *Journal of Experimental Psychology: General*, **147**(2), 170.
Levy, D.J., Glimcher, P.W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, **22**(6), 1027–38.
Lyubomirsky, S., King, L., Diener, E. (2005). The benefits of frequent positive affect: does happiness lead to success? *Psychological Bulletin*, **131**(6), 803.
McCullough, M.E., Kilpatrick, S.D., Emmons, R.A., Larson, D.B. (2001). Is gratitude a moral affect? *Psychological Bulletin*, **127**(2), 249.
Miller, E.K., Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**(1), 167–202.
Morishima, Y., Schunk, D., Bruhin, A., Ruff, C.C., Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, **75**(1), 73–9.
Nowak, M.A., Roch, S. (2007). Upstream reciprocity and the evolution of gratitude. *Proceedings of the Royal Society of London B: Biological Sciences*, **274**(1610), 605–10.
Nowak, M.A., Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, **437**(7063), 1291–8.
O'Doherty, J.P., Hampton, A., Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, **1104**(1), 35–53.
Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model. *Annual Review of Neuroscience*, **34**, 333–59.
Park, S.Q., Kahnt, T., Dogan, A., Strang, S., Fehr, E., Tobler, P.N. (2017). A neural link between generosity and happiness. *Nature Communications*, **8**, 15964.
Pfeiffer, T., Rutte, C., Killingback, T., Taborsky, M., Bonhoeffer, S. (2005). Evolution of cooperation by generalized reciprocity. *Proceedings of the Royal Society of London B: Biological Sciences*, **272**(1568), 1115–20.
Rand, D.G., Nowak, M.A. (2013). Human cooperation. *Trends in Cognitive Sciences*, **17**(8), 413–25.
Rangel, A., Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, **20**(2), 262–70.
Rankin, D.J., Taborsky, M. (2009). Assortment and the evolution of generalized reciprocity. *Evolution*, **63**(7), 1913–22.
Rilling, J.K., Sanfey, A.G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, **62**, 23–48.
Roalf, D. R. (2010). It's not fair! Behavioral and neural evidence that equity influences social economic decisions in healthy older adults. Scholar Archive. 516, https://digitalcommons.ohsu.edu/etd/516/?utm_source=digitalcommons.ohsu.edu%2Fetd%2F516&utm_medium=PDF&utm_campaign=PDFCoverPages.
Ruff, C.C., Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, **15**(8), 549–62.
Ruff, C.C., Ugazio, G., Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, **342**(6157), 482–4.
Rutte, C., Taborsky, M. (2007). Generalized reciprocity in rats. *PLoS Biology*, **5**(7), e196.
Sáez, I., Zhu, L., Set, E., Kayser, A., Hsu, M. (2015). Dopamine modulates egalitarian behavior in humans. *Current Biology*, **25**(7), 912–9.
Sanfey, A.G. (2007). Social decision-making: insights from game theory and neuroscience. *Science*, **318**(5850), 598–602.
Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, **300**(5626), 1755–8.
Schaafsma, S.M., Pfaff, D.W., Spunt, R.P., Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, **19**(2), 65–72.
Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, **42**, 9–34.
Seo, H., Lee, D. (2012). Neural basis of learning and preference during social decision-making. *Current Opinion in Neurobiology*, **22**(6), 990–5.
Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., Sack, A. T. (2014). Be nice if you have to–the neurobiological roots of strategic fairness. *Social cognitive and affective neuroscience*, **10**(6), 790–6.
Strang, S., Grote, X., Kuss, K., Park, S.Q., Weber, B. (2016). Generalized negative reciprocity in the dictator game—how to interrupt the chain of unfairness. *Scientific Reports*, **6**(1), 22316.
Strombach, T., Weber, B., Hangebrauk, Z., *et al.* (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences*, **112**(5), 1619–24.
Tricomi, E., Rangel, A., Camerer, C.F., O'Doherty, J.P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, **463**(7284), 1089–91.
Tusche, A., Böckler, A., Kanske, P., Trautwein, F.-M., Singer, T. (2016). Decoding the charitable brain: empathy, perspective taking, and attention shifts differentially predict altruistic giving. *Journal of Neuroscience*, **36**(17), 4719–32.
Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, **30**(3), 829–58.
Watanabe, T., Takezawa, M., Nakawake, Y., *et al.* (2014). Two distinct neural mechanisms underlying indirect reciprocity. *Proceedings of the National Academy of Sciences*, **111**(11), 3990–5.
Wu, Y., Zang, Y., Yuan, B., Tian, X. (2015). Neural correlates of decision making after unfair treatment. *Frontiers in Human Neuroscience*, **9**, 123.
Yu, R., Calder, A.J., Mobbs, D. (2014). Overlapping and distinct representations of advantageous and disadvantageous inequality. *Human Brain Mapping*, **35**(7), 3290–301.
Zhang, Y., Yu, H., Yin, Y., Zhou, X. (2016). Intention modulates the effect of punishment threat in norm enforcement via the lateral orbitofrontal cortex. *Journal of Neuroscience*, **36**(35), 9217–26.
Zhong, S., Chark, R., Hsu, M., Chew, S.H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *Neuroimage*, **129**, 95–104.