Intention Modulates the Effect of Punishment Threat in Norm Enforcement via the Lateral Orbitofrontal Cortex

Yuan Zhang,[1,2]* Hongbo Yu,[1]* Yunlu Yin,[1] and Xiaolin Zhou[1,3,4,5]

[1][illegible] 100871, [illegible] [2][illegible]
[illegible] 17456, [illegible] [3][illegible]
[illegible] [4][illegible] [5][illegible]
[illegible] 100871, [illegible]

Although economic theories suggest that punishment threat is crucial for maintaining social norms, counterexamples are noted in which punishment threat hinders norm compliance. Such discrepancy may arise from the intention behind the threat: unintentionally introduced punishment threat facilitates, whereas intentionally introduced punishment threat hinders, norm compliance. Here, we combined a dictator game and fMRI to investigate how intention modulates the effect of punishment threat on norm compliance and the neural substrates of this modulation. We also investigated whether this modulation can be influenced by brain stimulation. Human participants divided an amount of money between themselves and a partner. The partner (intentionally) or a computer program (unintentionally) decided to retain or waive the right to punish the participant upon selfish distribution. Compared with the unintentional condition, participants allocated more when the partner intentionally waived the power of punishment, but less when the partner retained such power. The right lateral orbitofrontal cortex (rLOFC) showed higher activation when the partner waived compared with when the computer waived or when the partner retained the power. The functional connectivity between the rLOFC and the brain network associated with intention/mentalizing processing was predictive of the allocation difference induced by intention. Moreover, inhibition or activation of the rLOFC by brain stimulation decreased or increased, respectively, the participants' reliance on the partner's intention during monetary allocation. These findings demonstrate that the perceived intention of punishment threat plays a crucial role in norm compliance and that the LOFC is casually involved in the implementation of intention-based cooperative decisions.

Key words: intention; lateral orbitofrontal cortex; norm compliance; punishment threat; tDCS

Significance Statement

Does punishment threat facilitate or hinder norm enforcement? So far, cognitive neuroscience research offers equivocal evidence. By directly manipulating the intention behind punishment threat, we demonstrate that intention modulates the effectiveness of punishment threat. Moreover, we show that inhibition or activation of the right lateral orbitofrontal cortex (rLOFC) decreased or increased the effect of punishment threat in the intentional context, but not in the unintentional context, suggesting the casual

Introduction

Social norms are widely shared rules about what constitutes appropriate behavior in social interactions (Bicchieri, 2006). Punishment is a ubiquitously adopted approach in human society to enforce norm compliance beyond the recipients voluntary action. Recent studies, however, provide divergent evidence concerning the effect of punishment threat on norm compliance. Studies reveal that participants achieve a higher level of norm compliance when punishment threat is present than when it is absent (Fehr and Gächter, 2002; Spitzer et al., 2007; Ruff et al., 2013). This is consistent with the deterrence theory, which holds that people are deterred from violating norms if they know the punishment will be severe (Carlsmith et al., 2002). Conversely,

Received Feb. 23, 2016; revised July 13, 2016; accepted July 16, 2016.

Author contributions: Y.Z., H.Y., and X.Z. designed research; Y.Z. and Y.Y. performed research; Y.Z., H.Y., and Y.Y. analyzed data; Y.Z., H.Y., Y.Y., and X.Z. wrote the paper.

This work was supported by the National Basic Research Program of China (973 Program Grant 2015CB856400) and the Natural Science Foundation of China (Grants 91232708 and 30110972). We thank Dr. Christian Ruff, Dr. Bolton K.H. Chau, Mr. Philip R. Blue, and two anonymous reviewers for their helpful comments and suggestions concerning previous versions of the manuscript.

The authors declare no competing financial interests.

*Y.Z. and H.Y. contributed equally to this work.

Correspondence should be addressed to Dr. Xiaolin Zhou, Department of Psychology, Peking University, 5 Yiheyuan Road, Beijing 100871 China. E-mail: xz104@pku.edu.cn

DOI:10.1523/JNEUROSCI.0595-16.2016

evidence also shows that punishment threat under certain circumstances hinders norm compliance. For example, in the trust game, the trustee returns less money to the investor when the investor imposes a punishment threat on the trustee (Fehr and Rockenbach, 2003; Gneezy and Rustichini, 2004; Houser et al., 2008; Li et al., 2009). The neural activity also shows contrasting patterns. Spitzer et al. (2007) found that activations in the lateral orbitofrontal cortex (LOFC) and dlPFC were positively correlated with individuals increase in norm compliance when punishment threat was present. In contrast, Li et al. (2009) observed decreased activations in the LOFC and ventromedial PFC (vmPFC) when punishment threat was present.

Closer examination of previous studies reveals that those reporting a detrimental effect typically adopted intentional punishment threat imposed by the interacting partner on behalf of his/her own interest (Fehr and Rockenbach, 2003; Li et al., 2009), whereas those reporting a facilitatory effect involved unintentional punishment threat, which was introduced by an impartial third-party (e.g., computer program) for the sake of fairness (Spitzer et al., 2007; Ruff et al., 2013). However, to our knowledge, no studies have investigated directly the role of intention behind punishment threat in norm enforcement. We hypothesized that the seemingly contradicting findings concerning the role of punishment threat could be reconciled if we take into account the intention behind the threat (Darley, 2009; Radke et al., 2012; Koster-Hale et al., 2013).

Of particular interest is the orbitofrontal cortex, a structure consistently implicated in computing social value and guiding social decision making (Rushworth et al., 2011; Rudebeck and Murray, 2014). We hypothesized that the LOFC may synthesize information about the presence of punishment threat and the intention by which it is imposed or forgone to form a unified signal that guides compliance behavior (Campbell-Meiklejohn et al., 2012).

To test our hypotheses, we manipulated the presence of punishment threat (Waive vs Retain) and the intention behind the threat (Intentional vs Unintentional) in a modified dictator game. By conducting an fMRI and two high-definition transcranial direct current stimulation (HD-tDCS) experiments, we examined the modulation of the neural processes of punishment threat by the intention behind such a threat. We were specifically interested in the role of the LOFC in mediating the influence of the perceived intention on norm compliance because this structure showed opposite effects when the threat was unintentional (Spitzer et al., 2007) or intentional (Li et al., 2009).

Materials and Methods

Participants

fMRI experiment. Thirty-five graduate and undergraduate students participated in the fMRI scanning. Ten were excluded (1 of them always transferred 0 yuan to the partner; 7 of them did not believe that they had interacted with different human partners, as indicated in the postexperiment manipulation check; 2 of them had excessive head movements $>3°$ in rotation or >3 mm in translation during the scanning), leaving 25 participants for data analysis (age range: 18 27 years, mean age: 21.2 years; 14 female). Due to technical problems, postscan questionnaire data were available for only 19 of these participants. We tested the robustness of online behavioral measures and postscan questionnaires

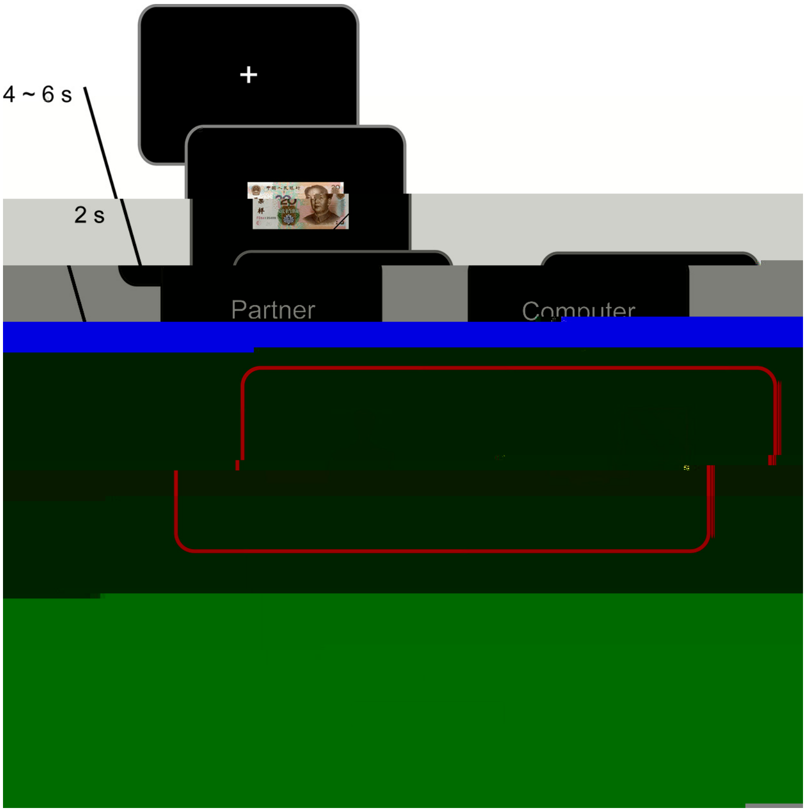

Figure 1. Procedure and task display. At the beginning of each trial, a picture of 20 yuan was presented, indicating the amount of money to be allocated. Then the phrases (in Chinese) "Partner decides" or "Computer decides" appeared on the screen, indicating the decider of the trial. This information remained on the screen for 2—5 s, followed by the decision outcome screen informing the participant about the actual decision for that trial. Then, a jittered fixation was presented, followed by a response bar on which the participant made the allocation. The critical event for fMRI data analysis was the outcome screen (red rectangle).

quence of the first block. Different pseudorandom sequences were created for different participants.

Before the fMRI scanning, we familiarized the participants with the task using a practice block consisting of eight trials. After the scanning, the participants were asked to rate their feelings of anger, disgust, hostility, trust, and gratitude under each condition on a 7-point Likert scale (1 = not at all, 7 = very). The participants were also asked to answer questions concerning the setup of the experiment (i.e., rating the intensity of perceived interaction with their partners and answering how many partners interacted with them during the game).

Neuroimaging data acquisition and preprocessing

Images were acquired using a GE Healthcare 3.0 T Medical Systems Discovery MR 750 with a standard head coil at Tongji University (Shanghai, China). T2*-weighted echoplanar images (EPI) were obtained with blood oxygenation level-dependent (BOLD) contrast. Forty transverse slices of 3 mm thickness that covered the whole brain were acquired in an interleaved order (repetition time = 2000 ms, echo time = 30 ms, field of view = 192 × 192 mm^2, flip angle = 90°).

The fMRI data were preprocessed and analyzed using Statistical Parametric Mapping software SPM8 (Wellcome Trust Department of Cognitive Neurology, London). Images were slice-time corrected, motion corrected, resampled to 3 mm × 3 mm × 3 mm isotropic voxels, normalized to MNI space, spatially smoothed with an 8 mm FWHM Gaussian filter, and temporally filtered using a high-pass filter with a cutoff frequency of 1/128 Hz.

Neuroimaging data analysis

Analysis of BOLD activation. Whole-brain analysis based on the general linear model was performed first at the participant level and then at the group level. At the participant level, we separately modeled the presentation of the regrouping cue (i.e., the fixation), the total amount cue, the cue of decisider (Computer or Partner), the cue of decision outcome (Retain or Waive), and distribution screen with a delta function convolved with a canonical hemodynamic response function (Friston et al., 1998). The decision outcome screen was separately modeled by four regressors, corresponding to the four possible combinations of two experimental factors (Decider × Threat). Further analyses of brain activation and connectivity were based on these regressors. Five regressors of no interest were included corresponding to the regrouping, the total amount cue, Decider (Computer vs Partner), and distribution. The number of button press was modeled as a covariate to rule out the potential confound of finger movement. The six rigid body parameters were also included to correct for the head motion artifact. We defined the contrasts corresponding to the interaction between decider and punishment threat: (Partner_Retain > Partner_Waive) > (Computer_Retain > Computer_Waive); the main effect of punishment threat: (Partner_Retain + Computer_Retain) > (Partner_Waive + Computer_Waive); and the effect of punishment threat under the intentional and unintentional contexts: Partner_Retain > Partner_Waive and Computer_Retain > Computer_Waive. The group-level analysis was performed using one-sample *t* test module implemented in the SPM8 software (random effect analysis). We focused on the interaction contrast because we hypothesized that norm compliance was modulated by the intention embedded in punishment threat, which was reflected in the interaction between Decider and Threat. The simple effects of punishment threat were included to further illustrate the pattern of interaction. To calculate the spatial extent criterion for the corrected cluster-level significance, we used the AlphaSim function implemented in REST, a software package publically available (Song et al., 2011). This function is based on the Monte Carlo simulation (http://afni.nimh.nih.gov/afni/doc/manual/AlphaSim). Using these tools, we determined that a corrected (familywise error, FWE) $p < 0.05$ cluster-level threshold was achieved with a minimum of 47 contiguous voxels, each significant at $p < 0.005$ (uncorrected).

We also performed independent region of interest (ROI) analyses based on the critical regions revealed in previous studies. Parameter estimates corresponding to the four critical regressors (i.e., Computer_Waive, Computer_Retain, Partner_Waive, Partner_Retain) were extracted from a cube (length of a side of the cube was 3 voxels) containing 27 voxels around the following coordinates, including the rLOFC ([44, 42, −6]; Spitzer et al., 2007) and the vmPFC ([4, 56, −4]; Li et al., 2009). These ROIs were chosen because they were the core structures for norm compliance identified in the previous studies.

Psychophysiological interaction (PPI). Given that the vmPFC and the LOFC were found to play an important role in mediating the relationship between the processing of intention behind punishment threat and norm compliance behavior (see Results), a further question that naturally follows is: from which brain regions does the information concerning other's intention come? If our hypothesis is correct (i.e., that the responses of vmPFC and LOFC to punishment threat are modulated by the processing of intention), then we should observe functional interplay between the vmPFC and LOFC, on the one hand, and the typical intention processing/mentalizing network (e.g., dorsomedial prefrontal cortex, dmPFC; temporoparietal junction, TPJ; precuneus; Molenberghs et al., 2016), on the other hand. To test this conjecture, we performed a PPI analysis (Friston et al., 1997) using the vmPFC and the LOFC identified in our whole-brain analysis as seed regions. We calculated PPI maps corresponding to the contrast Partner_Retain Computer_Retain (i.e., intentional punishment threat hinders norm compliance) and Partner_Waive Computer_Waive (i.e., refraining from the threat of punishment facilitates norm compliance). To test the possibility that the strength of such functional connectivity is modulated by individuals' susceptibility to the intention effect, we added the difference in allocation corresponding to each of these contrasts as a group-level covariate. We then used the one-sample t test in SPM8 to perform statistical analysis. The statistic threshold was the same as indicated above.

Brain stimulation experiment

To test the causal role of the rLOFC in mediating the influence of intention on punishment threat, we performed two brain stimulation experiments using HD-tDCS. The first group of participants (n 22) received cathodal stimulation and sham stimulation in two experiment sessions. Half of the participants received cathodal stimulation over the rLOFC in the first experiment day and received sham stimulation over the same area in the second experiment day. The other half of the participants received the reversed stimulation protocol. The second group of participants (n 20) received anodal stimulation and sham stimulation in two experiment sessions. Similar to the cathodal experiment, half of these participants received anodal stimulation over the rLOFC in the first experiment day and received sham stimulation over the same area in the second experiment day. The other half of the participants received the reversed stimulation protocol. Therefore, both of the two HD-tDCS experiments used a within-participant design; moreover, to avoid carry-over effects of brain stimulation, sessions were separated by at least 24 h for each participant. The behavioral protocol was identical to the fMRI experiment.

Figure 2. Behavioral results. A, Monetary allocation. B, Self-reported feeling of being trusted. C, D, Patterns of self-reported positive (happiness, benevolence, gratitude) and negative affect (sadness, anger, fear, hostility, aversi

HD stimulation was delivered using a multichannel stimulation adapter (Soterix Medical, 4 1, Model C3) connected to the constant current stimulator (Soterix Medical, Model 1300-A). A 41 montage consisting of five sintered Ag/AgCl ring electrodes was used and these electrodes were arranged on the skull in a 4 1 ring configuration as suggested by the previous literature (Minhas et al., 2010). The electrodes were held in place in plastic electrode holders in a fitted cap (EASYCAP). The electrode holders were filled with SignaGel, creating a gel contact of 4 cm^2 per electrode. The position of the electrode was identified and adjusted using HD-Explore software (Soterix Medical), which uses a finite-element-method modeling approach to quantify electric field intensity throughout the brain (Datta et al., 2009; Dmochowski et al., 2011; Kempe et al., 2014). The locations of the electrodes were chosen by selecting the 10–20 EEG sites that would optimally target the rLOFC in our fMRI study. Therefore, we selected central electrode as FP2 in the 10–20 EEG coordinate system and surrounded it with three return electrodes at F2, F8, Fp1, and one return electrode at the lower eyelid (each at a distance of 6 cm from the central electrode). For active anodal/cathodal stimulation, participants received a constant current of 2.0 mA for 20 min. Stimulation started 8 min before the task and was delivered during the entire course of the task (20 min), with an additional 30 s ramp-up at the beginning of stimulation and 30 s ramp-down at the end. For the sham stimulation, the initial 30 s ramp-up was immediately followed by the 30 s ramp-down and there was no stimulation for the rest of the session. For both the experimental and sham stimulation conditions, participants felt a little uncomfortable initially, but were unaware of stimulation before the task started.

Compared with the classic conventional bipolar tDCS, HD-tDCS has been shown to have better spatial focality, larger effect on cortical excitability, and longer after effects (Datta et al., 2009; Caparelli-Daquer et al., 2012; Kuo et al., 2013). Although HD-tDCS is associated with stronger scalp sensations than conventional tDCS, it has been shown to be safe and tolerable with applications of up to 2.0 mA for 20 min (Minhas et al., 2010; Borckardt et al., 2012; Kuo et al., 2013).

Results

Behavioral results

To determine whether the participants' allocation was jointly modulated by the presence of threat and the intention behind it, we performed a Decider (Computer vs Partner) by Threat (Waive vs Retain) repeated-measures ANOVA for the allocation in the fMRI experiment. The only significant effect was the interaction between Decider and Threat ($F_{(1,24)}$ 27.15, p 0.001; Fig. 2A). Pairwise comparison showed that, compared with the corresponding unintentional conditions (i.e., the Computer as the decider), the participants allocated more to the partner when the partner intentionally waived ($F_{(1,24)}$ 13.43, p 0.001) and less when the partner intentionally retained the punishment threat ($F_{(1,24)}$ 8.07, p 0.005). Moreover, compared with the condition in which the partner intentionally retained the punishment threat (i.e., Partner_Retain), the participants allocated more to the partner in the condition in which the partner voluntarily waived the punishment threat (Partner_Waive) ($F_{(1,24)}$ 4.39, p 0.05). The same pattern of interaction was observed in the behavioral validation experiment ($F_{(1,23)} = 10.83$, $p < 0.001$). Pairwise comparison showed that, compared with the Computer_Waive condition, participants allocated significantly more to the partner in the Partner_Waive condition ($F_{(1,23)} = 4.85$, $p < 0.05$); compared with the Computer_Retain condition, participants allocated less to the partner in the Partner_Retain condition ($F_{(1,23)} = 3.33$, $p = 0.081$).

For the emotional ratings (Fig. 2*B*, *D*), we averaged the ratings of happiness, benevolence, and gratitude to form an indicator of positive affect and the ratings of sadness, anger, fear, aversion, and hostility to form an indicator of negative affect. We then performed a repeated-measures ANOVA with emotional valence (Positive vs Negative), Decider (Partner vs Computer), and Threat (Retain vs Waive) as within-participant factors. Note that we only had the postscan questionnaire data for 19 of the 25 fMRI participants. The three-way interaction was significant ($F_{(1,18)} = 20.58$, $p < 0.001$). We then performed two two-way repeated-measure ANOVAs separately for the positive and negative affect indicators. For the positive affect, the two-way interaction was significant ($F_{(1,18)} = 28.94$, $p < 0.001$). Pairwise comparison showed that the positive affect was higher in the Partner_Waive condition than in the Computer_Waive and the Partner_Retain conditions ($F > 37$, $p < 0.001$). For the negative affect, the two-way interaction was significant ($F_{(1,18)} = 7.12$, $p < 0.05$). The negative affect was higher in the Partner_Retain condition than in the Computer_Retain and the Partner_Waive conditions ($F > 5$, $p < 0.05$). Moreover, we performed a two-way ANOVA on the ratings of perceived trust. The interaction was significant ($F_{(1,18)} = 33.52$, $p < 0.001$). Pairwise comparison showed that the perceived trust was higher in the Partner_Waive condition than in the Computer_Waive condition ($F_{(1,18)} = 68.16$, $p < 0.00$) and the Partner_Retain condition ($F_{(1,18)} = 32.03$, $p < 0.001$).

Again, the postexperiment ratings of behavioral validation experiment replicated the behavioral data of the fMRI experiment. For positive emotions, the Decider-by-Threat interaction was significant ($F_{(1,23)} = 49.79$, $p < 0.001$). Pairwise comparison showed that positive affect was higher in the Partner_Waive condition than in the Computer_Waive and the Partner_Retain conditions ($F > 73$, $p < 0.001$). For the negative affect, the two-way interaction was marginally significant ($F_{(1,23)} = 3.80$, $p = 0.064$). The negative affect was higher in the Partner_Retain condition than in the Computer_Retain and the Partner_Waive conditions ($F > 11$, $p < 0.01$). For perceived trust, the Decider-by-Threat interaction was significant ($F_{(1,23)} = 22.70$, $p < 0.001$). The perceived trust was higher in the Partner_Waive condition than in the Computer_Waive condition ($F_{(1,23)} = 52.18$, $p < 0.001$) and the Partner_Retain condition ($F_{(1,23)} > 27.14$, $p < 0.001$). Together, these results strongly indicate that intentionally introducing punishment threat elicits strong negative emotions, whereas intentionally waiving punishment threat elicits strong positive emotions such as gratitude and the feeling of being trusted.

Whole-brain analysis of the neuroimaging data

When the decision was to retain the punishment threat, the participants were facing certain danger and provocation regardless of whether it was made by the partner or by the computer program. Previous studies have shown that several brain areas related to mentalizing (e.g., dmPFC, TPJ) and affective salience (e.g., thalamus, insula, caudate) are recruited in situations of reactive aggression and hostility (Krämer et al., 2007, 2011; Beyer et al., 2015). Consistent with these findings, the main effect contrast Retain > Waive revealed activations in the dmPFC, thalamus, dorsal caudate, and TPJ (Fig. 3*A*).

To test our hypothesis concerning the modulation of intention on the effect of punishment threat, we examined the interaction contrast (Partner_Waive > Computer_Waive) > (Partner_Retain > Computer_Retain). This contrast revealed activations in the bilateral LOFC (left LOFC: MNI coordinates = [−42, 32, 1], cluster size = 77, $t_{(24)} = 3.66$; rLOFC: MNI coordinates = [42, 35, −5], cluster size = 72, $t_{(24)} = 3.85$; Fig. 3*B*). Given that we did not observe an interaction in the vmPFC at the current threshold level, we performed an ROI-based analysis within a predefined vmPFC ROI (small volume correction within an 8-mm-radius sphere around [4, 56, −4], the coordinates reported in Li et al., 2009). This analysis did reveal a significantly activated cluster (MNI coordinates = [3, 56, −8]; cluster size = 14; $t_{(24)} = 3.32$; peak-level $p_{\mathrm{FWE}} < 0.05$; Fig. 3*B*). The reversed contrast did not reveal any significant clusters.

To illustrate the interaction more clearly, we decomposed the interaction into two separate contrasts: Computer_Retain > Computer_Waive, which corresponded to unintentional punishment threat (Spitzer et al., 2007), and Partner_Waive > Partner_Retain, which corresponded to intentionally withdrawing the punishment right (Li et al., 2009). The former contrast (Fig. 3*C*) revealed activation clusters in the left LOFC (MNI coordinates = [−39, 32, 1], cluster size = 103, $t_{(24)} = 4.18$) and the left caudate (MNI coordinate = [−9, 8, 1], cluster size = 106, $t_{(24)} = 3.70$). The latter contrast (Fig. 3*D*) revealed only one activation cluster in the rLOFC (MNI coordinate = [39, 35, −5], cluster size = 48, $t_{(24)} = 3.88$).

ROI-based analysis of the neuroimaging data

To buttress the findings derived from the whole-brain analysis, we performed further analyses for predefined ROIs: the vmPFC and the LOFC. We hypothesized that, if vmPFC activation reflected positive social value (eg, mutual trust) perceived in the dyadic interaction, then it should show higher activation when the partner intentionally waived the punishment threat, an action that may convey trust (Fig. 2*B*), than when the partner retained the threat. To test this hypothesis, we performed a small volume correction within the vmPFC ROI (8 mm-radius sphere around [4, 56, −4], coordinates reported in Li et al., 2009). This analysis revealed a significantly activated cluster in the vmPFC ROI (MNI coordinates = [3, 56, −8]; cluster size = 17; $t_{(24)} = 3.41$; peak-level $p_{\mathrm{FWE}} = 0.013$; Fig. 3*D*). Concerning the rLOFC, we hypothesized that its responses to punishment threat should be modulated by the intentionality behind the threat. Specifically, the rLOFC activation should be higher in the Computer_Retain condition than in the Computer_Waive condition, whereas the opposite pattern should be observed for the Partner conditions. To this end, we performed a small volume correction within the rLOFC ROI (8-mm-radius sphere around [44, 42, −6], coordinates reported in Spitzer et al., 2007). Within this rLOFC ROI, the contrast Computer_Retain > Computer_Waive revealed a significantly activated cluster centered around the MNI coordinates [51, 38, −2] (cluster size = 2; $t_{(24)} = 2.91$; peak-level $p_{\mathrm{FWE}} < 0.05$), while the contrast Partner_Waive > Partner_Retain revealed a significantly activated cluster centered around the MNI coordinates [39, 38, −5] (cluster size = 15; $t_{(24)} = 3.54$; peak-level $p_{\mathrm{FWE}} < 0.01$). Such dissociation confirmed our hypothesis concerning the rLOFC.

Moreover, the parameter estimates extracted from the predefined rLOFC and vmPFC ROIs (27 voxels around the coordinates reported in Spitzer et al., 2007 and Li et al., 2009 for rLOFC and vmPFC, respectively) exhibited a pattern generally consistent with our findings derived from the small volume correction analysis (Fig. 3E,F). We performed repeated-measures ANOVAs on the parameter estimates and report the statistical details in Table 1. The Decider-by-Threat interaction was significant for both the rLOFC and the vmPFC. Specifically, for the vmPFC, the activation was significantly higher in the Partner_Waive condition than in the Partner_Retain condition (i.e., the same as reported in Li et al., 2009) and was also significantly higher than in the Computer_Waive condition, consistent with the social value representation view of vmPFC function (Ruff and Fehr, 2014). For the rLOFC, the parameter estimates appeared to be higher in the Partner_Waive condition than in the Partner_Retain condition and the parameter estimates appeared to be higher in the Computer_Retain condition than in the Computer_Waive condition, although these differences did not reach statistical significance.

Functional connectivity (PPI) analysis

We performed PPI analyses to test whether the functional connectivity between the mentalizing network and the left vmPFC or the rLOFC was modulated by experimental manipulation and whether such connectivity was predictive of participants norm compliance behavior. The functional connectivity (for the contrast Partner_Waive > Computer_Waive) between the rLOFC and several brain areas in the typical mentalizing network (e.g., dmPFC, TPJ, and precuneus) was positively correlated with the difference in allocation amount between the Partner_Waive and Computer_Waive conditions (Fig. 4, yellow areas; Table 2). Similarly, the functional connectivity (for the contrast Partner_Retain > Computer_Retain) between the rLOFC and several brain areas in the typical mentalizing network (e.g., dmPFC, TPJ, and precuneus) was positively correlated with the difference in allocation amount between the Computer_Retain and Partner_Retain conditions (Fig. 4, blue areas; Table 2). No significant result was revealed by the PPI analysis with vmPFC.

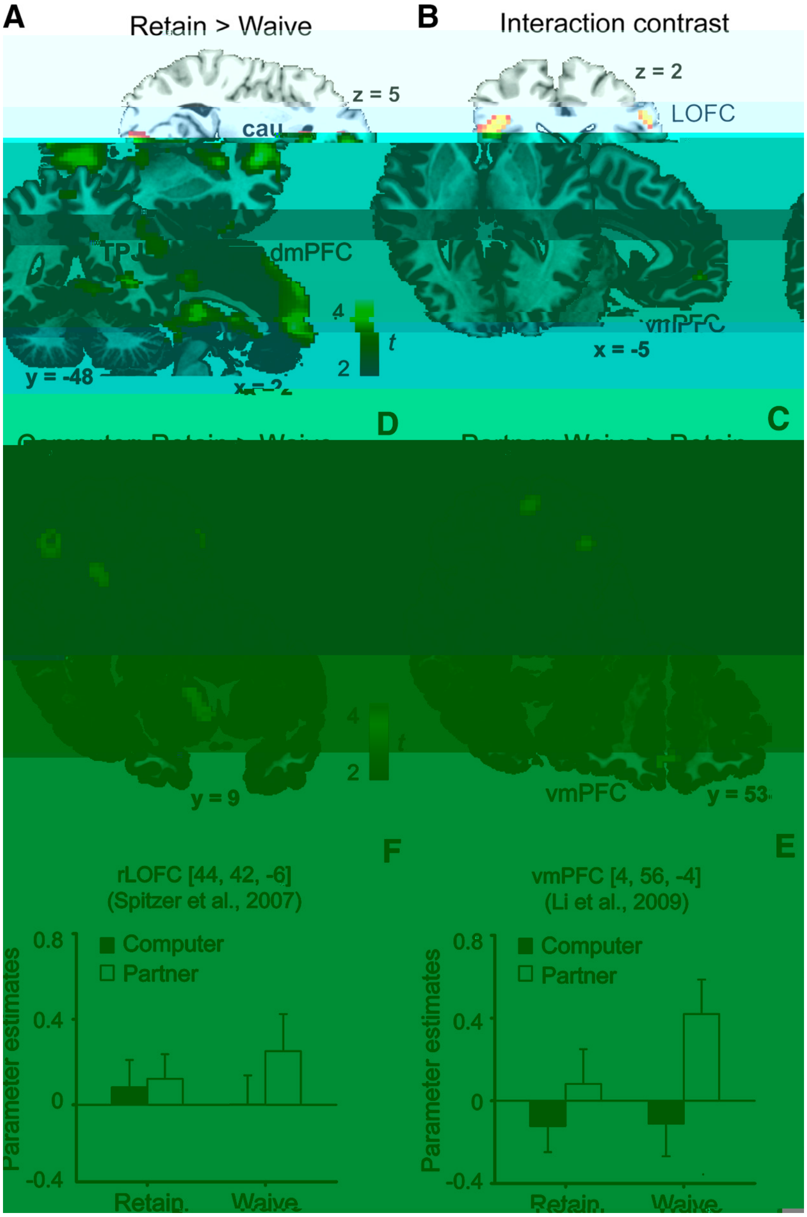

Figure 3. Analysis of brain activation. A, The whole-brain main effect contrast Retain > Waive revealed activation in the areas typically associated with intentional/mentalizing processing (e.g., dmPFC, TPJ) and affe thalamus, dorsal caudate). B, The whole-brain interaction contrast (Partner_Waive > Computer_Waive) > (Partner_Retain > Computer_Retain) revealed activation in the bilateral LOFC. C, The contrast Computer_Retain > Computer_Waive revealed activation in the bilateral LOFC and the left caudate. D, The contrast Partner_Waive > Partner_Retain revealed activation in the rLOFC and the vmPFC. E, F, ROI analyses on the activation in the rLOF (Spitzer et al., 2007) and the vmPFC (Li et al., 2009) based on the previous literature. No activation was found for P Retain > Waive at the current threshold. Detailed statistical results are provided in Table 1. Error bars indicate SE.

Brain stimulation (HD-tDCS) results

For each of the tDCS experiments, we performed a repeated-measures ANOVA with Stimulation Type (Cathodal/Anodal vs Sham), Decider (Computer vs Partner), and threat (Retain vs Waive) as within-participant factors. For the cathodal experiment, the three-way interaction was significant ($F_{(1,21)} = 5.97$, $p < 0.05$; Fig. 5A). We then performed a two-way ANOVA focusing on the data in which the partner determined the presence or absence of the punishment threat. The interaction between Stimulation Type and Threat was significant ($F_{(1,21)} = 11.10$, $p <$

Table 1. ROI analysis of brain activations

Contrast	rLOFC $F_{(1,24)}$	rLOFC p	vmPFC $F_{(1,24)}$	vmPFC p
Interaction	4.99	0.035	7.73	0.010
Partner_Waive vs Partner_Retain	2.41	0.134	4.51	0
Computer_Waive vs Computer_Retain	2.11	0.159	0.037	
Partner_Waive vs Computer_Waive	7.99	0.009	15.43	
Partner_Retain vs Computer_Retain	0.27	0.605	2.63	

rLOFC, right lateral orbitofrontal cortex; vmPFC, ventromedial prefrontal cortex.

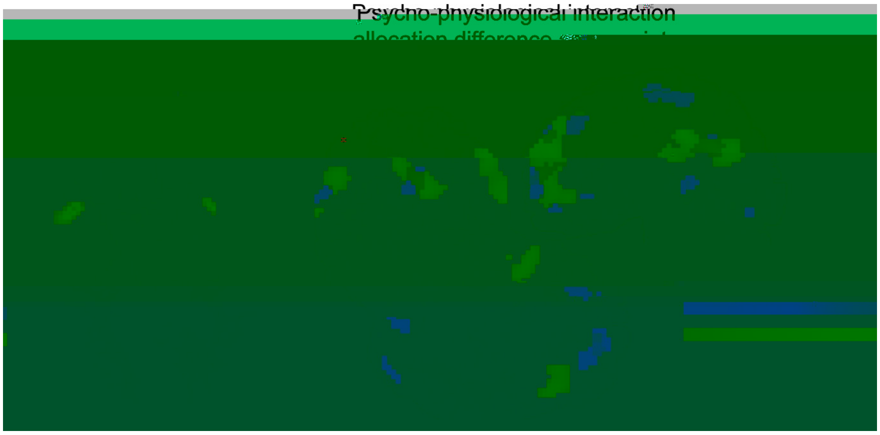

Figure 4. Results of the PPI analysis. The rLOFC identified in the whole-brain contrast was used as the seed region. The contrast Partner_Retain > Computer_Retain and Partner_Waive > Computer_Waive, with the allocation differences as covariate, revealed a series of brain areas overlapping with the mentalizing network. The functional connectivity (for the contrast Partner_Retain > Computer_Retain) between the rLOFC and the revealed brain areas (blue areas) positively correlated with the difference in allocation amount between the Computer_Retain and Partner_Retain conditions. Similarly, the functional connectivity (for the contrast Partner_Waive > Computer_Waive) between the rLOFC and the yellow areas positively correlated with the difference in allocation amount between the Partner_Waive and Computer_Waive conditions.

Table 2. Brain activations revealed by the PPI covariate contrast ($p < 0.005$ uncorrected at voxel level, cluster level $p < 0.05$, FWE corrected)

Regions	Hemi	Max *T*-value	Cluster size (voxels)	MNI coordinates		
				x	*y*	*z*
Partner_Waive > Computer_Waive[a]						
dmPFC	L/R	5.83	1651		12	41 46
dlPFC	L	5.53	178	−36	11	43
	R	4.79	136	57	14	37
Insula	L	4.79	149	−30	14	−14
	R	5.35	197	45	17	−14
Precuneus	L/R	5.14	856	3	−67	40
Angular	L	4.42	246	−51	−58	31
	R	5.07	285	48	−64	40
Partner_Retain > Computer_Retain[b]						
dmPFC	L/R	6.26	1400		6	62 10
LOFC	L	4.11	48	−51	17	1
SFG	L	5.04	383	−42	14	40
Putamen	L	4.26	163	−24	14	13
STS	R	4.35	70	66	−10	−2
Precuneus	L/R	5.53	511	3	−55	31
Angular	L	4.81	496	−45	−49	28

dlPFC, dorsolateral prefrontal cortex; dmPFC, dorsomedial prefrontal cortex; LOFC, lateral orbitofrontal cortex; SFG, superior frontal gyrus; STS, superior temporal sulcus.

[a] Positive correlation with allocation difference (Partner_Waive > Computer_Waive).

[b] Positive correlation with allocation difference (Computer_Retain > Partner_Retain).

0.005). Pairwise comparison showed that, relative to the sham stimulation, the cathodal stimulation decreased the participants allocation when the partner s decision was to waive the punishment threat ($F_{(1,21)} = 4.91$, $p < 0.05$) and increased the allocation when the partner s decision was to retain the punishment threat ($F_{(1,21)} = 5.56$, $p < 0.05$). The same analysis was also applied to the Computer conditions, but neither the main effect nor the interaction was significant.

For the anodal experiment, the three-way interaction was significant ($F_{(1,19)} = 6.00$, $p < 0.05$; Fig. 5B). We then performed a two-way ANOVA focusing on the Partner conditions. The interaction between Stimulation Type and Threat was significant ($F_{(1,19)} = 20.68$, $p < 0.001$). Pairwise comparison showed that, relative to the sham stimulation, the anodal stimulation increased the participants allocation when the partner s decision was to waive the punishment threat ($F_{(1,19)} = 8.87$, $p < 0.01$) and decreased the allocation when the partner s decision was to retain the punishment threat ($F_{(1,19)} = 13.57$, $p < 0.005$). The same analysis applied to the Computer conditions revealed neither a significant main effect nor a significant interaction.

To better illustrate and examine the effects of brain stimulation (both inhibition and activation) on intentional/unintentional norm enforcement, we calculated the effect of punishment threat (i.e., the amount transferred in the Waive condition minus the amount transferred in the Retain condition) in the intentional (Partner) and unintentional (Computer) contexts for both the cathodal and anodal groups (Fig. 5C). We then performed two repeated-measures ANOVAs with Stimulation Type (Cathodal/Anodal vs sham) and Decider (Computer vs Partner) as within-participant factors. For the cathodal group, the interaction between Stimulation Type and Threat was significant ($F_{(1,21)} = 5.96$, $p < 0.05$). Relative to the sham stimulation, the cathodal stimulation decreased the effect of punishment threat mainly in the intentional context ($F_{(1,21)} = 11.10$, $p < 0.005$), but not in the unintentional context ($F_{(1,21)} = 3.60$, $p = 0.072$). For the anodal group, the interaction between stimulation type and threat was significant ($F_{(1,19)} = 5.99$, $p < 0.05$). Relative to the sham stimulation, the anodal stimulation increased the effect of punishment threat only in the intentional context ($F_{(1,19)} = 20.68$, $p < 0.001$), not in the unintentional context ($F_{(1,19)} < 1$, $p > 0.1$).

Two features of this pattern are worth noting. First, inhibition and activation of the rLOFC had opposite effects on the participants norm compliance behavior (i.e., monetary allocation): whereas activation of this area tended to increase the effect of waiving the punishment threat on norm compliance (cf. filled and empty red dots in Fig. 5C), inhibition of this area tended to decrease this effect (cf. filled and empty blue diamonds in Fig. 5C). Second, the brain stimulation took effect mainly in the intentional context (cf. difference between filled-empty pairs on the Partner side with its counterparts on the Computer side in Fig. 5C).

Discussion

Our behavioral results demonstrated that the perceived intention modulates the effect of punishment threat on norm compliance. Specifically, we observed a detrimental effect of punishment threat in the intentional context (i.e., partner as decider), consistent with previous studies (Fehr and Rockenbach, 2003; Gneezy and Rustichini, 2004; Li et al., 2009). In the unintentional context (i.e., computer as decider), although we did not observe a facilitatory effect of punishment threat, as previous studies did (Fehr and Gächter, 2002; Spitzer et al., 2007; Ruff et al., 2013), the disappearance of the detrimental effect suggests that intention does play an important role in the effectiveness of punishment threat.

The intention underlying punishment threat may influence a key factor in norm compliance behavior: the perceived legitimacy of authority. When an impartial computer program or a third party decides to retain the power to punish the allocator, it is conceived that the retention of punishment threat is on behalf of the social norms themselves. This argument is supported by both our study, which revealed no detrimental effects on norm compliance, and previous studies, which revealed facilitatory effects on norm compliance (Spitzer et al., 2007; Ruff et al., 2013). In contrast, when the partner (i.e., the second party), whose interest is directly affected by the allocation, decides to retain the power to punish the allocator, the purpose of the punishment threat is dubious. It may be perceived, not as a way to maintain justice, but rather as a way to serve selfish interest or to signal distrust, resulting in reduced norm compliance (Dickinson and Villeval, 2008). This argument is supported by our behavioral results and the emotion self-reports indicating that intentional retention of punishment threat elicits stronger negative feelings and less amount of allocation than unintentional retention or intentional waiving of punishment threat. In addition, intention can function in, not only a negative way, but also a positive way. We found that, compared with both unintentional waiving and intentional retention of punishment threat, participants reported stronger positive feelings (e.g., being trusted, more grateful) and allocated more to the partner when the latter intentionally waived the power to punish the former.

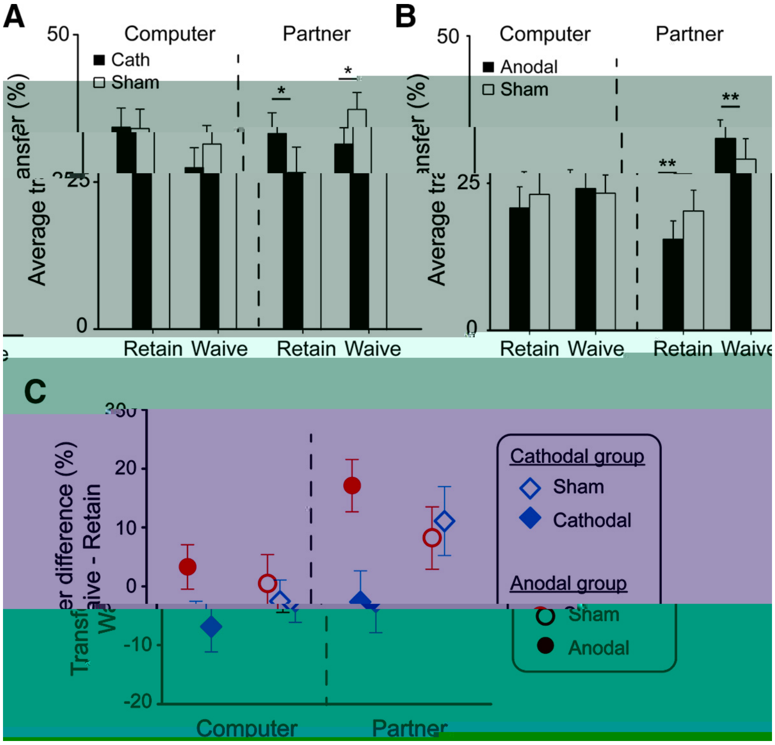

Figure 5. Results of the HD-tDCS experiments. The allocation as the function of Stimulation Type (A Sham), Decider (Computer vs Partner), and Threat (Retain vs Waive) in the cathodal (*A*) and anodal (*B*) experiments. *C*, Cathodal stimulation decreased, whereas anodal stimulation increased, the effect of punishment th transferred in the Waive condition minus the amount transferred in the Retain condition) in the int context. Error bars indicate SE. $^{*}p < 0.05$; $^{**}p < 0.01$.

Houser et al. (2008) also manipulated intention but did not find any effect of intention on norm compliance. The discrepancy between their findings and ours may come from two sources. First, intention was a within-participant factor in our study, but a between-participant factor in their study. Therefore, participants who experienced both intentional and unintentional contexts may exhibit a strengthened contrast between the two contexts, which amplifies the difference between intentional and unintentional punishment threat on the perceived legitimacy of authority. Second, the partner s demand of the allocation portion was not revealed in our study, but was revealed in Houser et al. (2008). Because the participants clearly knew their partner s demand in Houser et al. (2008), they could easily calculate all of the outcomes (i.e., outcome when keeping the entire investment and being punished vs outcome when returning what the partner demanded) and select the most profitable strategy. Such an experimental setup may drive participants to utility-driven strategies, crowding out the influence of intention.

The average transfer in our study was between 30% and 40% of the endowed amount, even in the punishment threat conditions. This was relatively low compared with previous studies, which usually reported 40% average transfer (Spitzer et al., 2007) or 40 50% transfer (Ruff et al., 2013) under punishment threat. The discrepancy may be due to the intensity of punishment threat. In the current study, the intensity was relatively low (4 yuan; the whole allocation endowment was 20 yuan) compared with the previous studies. The intensity of punishment threat can modulate its effect on norm enforcement (Gneezy and Rustichini, 2004) and, intuitively, when the punishment threat is large enough, it will dominate people s consideration about norm compliance behavior. The discrepancy between the studies, however, does not eliminate the validity of the intention effect that we observed at small amounts of punishment threat. As Gneezy and Rustichini (2004) noted, we have no evidence to support the hypothesis that the psychological and behavioral factors that drive the reaction to small fines or rewards disappear completely when higher amounts are offered or charged, thus reducing the explanation of behavior to a choice of the most convenient combination of effort and reward.

Of particular interest to us is the LOFC, which has been consistently implicated in norm compliance, but has showed opposite activation patterns depending on whether punishment threat was introduced intentionally or unintentionally (Spitzer et al., 2007; Li et al., 2009). Some propose that the LOFC functions to encode the punishment threat based on the findings that higher LOFC activation is associated with more norm compliance behaviors under (unintentional) punishment threat (Spitzer et al., 2007). Our results indicated that this could not be the whole story because the LOFC also showed higher activation when the partner intentionally waived the punishment threat. An alternative interpretation, which fits better with both the previous and the current findings, is that the LOFC integrates information from various sources (e.g., intention, emotion, material interest, etc.) and outputs a decision as to whether to conform to the social norm (Rolls and Grabenhorst, 2008). When the presence or absence of the punishment threat is determined by a nonintentional computer program, it is possible that the decision to conform is dominated by the consideration of material interests; that is, the rational calculation of gains and losses. This argument is supported by findings in the current study and Spitzer et al. (2007)

that the norm compliance behavior and LOFC activation were higher in the presence of punishment threat. When the presence or absence of punishment threat is determined by the partner, it conveys important social information, such as trust or distrust. In such contexts, the LOFC and the participant s norm compliance are sensitive to the social signal behind the punishment threat. This conjecture was buttressed by our brain stimulation data: inhibition or activation of the rLOFC by tDCS decreased or increased the effect of partner s intention on norm compliance behavior. Note that we do not claim the laterality of LOFC because we do not have an *a priori* hypothesis. We focused our analysis on the right rather than the left LOFC because the discrepancy between Spitzer et al. (2007) and Li et al. (2009) was on the rLOFC. As can be seen from Figure 3, *B* *D*, although both the left and right LOFC were revealed in the interaction contrast, only the rLOFC was activated in both simple effect contrasts: Computer_Retain > Computer_Waive and Partner_Waive > Partner_Retain.

The brain stimulation took effect mainly in the intentional context, not in the unintentional context, suggesting that the inhibition or activation of the rLOFC may not affect its function in punishment threat processing, but may disrupt or facilitate its function in interacting with other brain regions that could provide social information (e.g., intention, emotion). This argument was supported by our results showing that the functional connectivity between the rLOFC and the brain network typically associated with intention/mentalizing processing (including dmPFC, TPJ, and precuneus; Molenberghs et al., 2016) was predictive of the effect of intention on norm compliance. Moreover, the functional connectivity (Partner_Waive > Computer_Waive) between the bilateral insula and the rLOFC positively correlated with the increase in norm compliance behavior. The bilateral insula was found to be associated with the aversion of anticipated guilt by not honoring others trust (Chang et al., 2011), which may drive individuals to conform to social norms and to show mutual respect in social interaction (Charness and Dufwenberg, 2006). Therefore, it is conceivable that the insula encodes the potential guilt that could arise if the participant fails to honor the partner s trust and benevolence (e.g., in the Partner_Waive condition). Such emotional information may be projected to the LOFC to bias the participants norm compliance behavior.

Finally, we also found higher activation in the vmPFC when the partner waived the power to punish the participant compared with when the partner retained or when the computer waived such power. This is consistent with Li et al. (2009), in which the vmPFC showed higher activation when the partner voluntarily waived the power to punish the participants. Ample evidence has implicated the vmPFC in computing both social and nonsocial reward values (Haber and Knutson, 2010; Bartra et al., 2013; Ruff et al., 2014). For example, the act of saving money is valued differently and elicits differential activation in the vmPFC according to whether the saving is for charitable donation (higher social value) or for self-interest (lower social value) (Cooper et al., 2010; Hare et al., 2010). We argue that the partner s voluntary waiving of the power to punish (i.e., trust and benevolence) is perceived to be most valuable to the individuals.

In conclusion, by combining an interactive game, fMRI, and HD-tDCS, we demonstrate that intention plays an important role in the effectiveness of punishment threat on norm compliance and that the LOFC is casually involved in the implementation of intention-based cooperative decisions.

References

Bartra O, McGuire JT, Kable JW (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76:412 427. CrossRef Medline

Beyer F, Munte TF, Gottlich M, Kramer UM (2015) Orbitofrontal cortex reactivity to angry facial expression in a social interaction correlates with aggressive behavior. Cereb Cortex 25:3057 3063. CrossRef Medline

Bicchieri C (2006) The grammar of society: the nature and dynamics of social norms. New York: Cambridge University.

Borckardt JJ, Bikson M, Frohman H, Reeves ST, Datta A, Bansal V, Madan A, Barth K, George MS (2012) A pilot study of the tolerability and effects of high-definition transcranial direct current stimulation (HD-tDCS) on pain perception. J Pain 13:112 120. CrossRef Medline ments from neural representations of intentions. Proc Natl Acad Sci U S A 110:5648–5653. CrossRef Medline

Krämer UM, Jansma H, Tempelmann C, Münte TF (2007) Tit-for-tat: the neural basis of reactive aggression. Neuroimage 38:203–211. CrossRef Medline

Krämer UM, Riba J, Richter S, Münte TF (2011) An fMRI study on the role of serotonin in reactive aggression. PLoS One 6:e27668. CrossRef Medline

Kuo HI, Bikson M, Datta A, Minhas P, Paulus W, Kuo MF, Nitsche MA (2013) Comparing cortical plasticity induced by conventional and high-definition 4×1 ring tDCS: a neurophysiological study. Brain Stimul 6:644–648. CrossRef Medline

Li J, Xiao E, Houser D, Montague PR (2009) Neural responses to sanction threats in two-party economic exchange. Proc Natl Acad Sci U S A 106:16835–16840. CrossRef Medline

Minhas P, Bansal V, Patel J, Ho JS, Diaz J, Datta A, Bikson M (2010) Electrodes for high-definition transcutaneous DC stimulation for applications in drug delivery and electrotherapy, including tDCS. J Neurosci Methods 190:188–197. CrossRef Medline

Molenberghs P, Johnson H, Henry JD, Mattingley JB (2016) Understanding the minds of others: A neuroimaging meta-analysis. Neurosci Biobehav Rev 65:276–291. CrossRef Medline

Radke S, Güroğlu B, de Bruijn ER (2012) There's something about a fair split: intentionality moderates context-based fairness considerations in social decision making. PLoS One 7:e31491. CrossRef Medline

Rolls ET, Grabenhorst F (2008) The orbitofrontal cortex and beyond: from affect to decision-making. Prog Neurobiol 86:216–244. CrossRef Medline

Rudebeck PH, Murray EA (2014) The orbitofrontal oracle: cortical mechanisms for the prediction and evaluation of specific behavioral outcomes. Neuron 84:1143–1156. CrossRef Medline

Ruff CC, Fehr E (2014) The neurobiology of rewards and values in social decision making. Nat Rev Neurosci 15:549–562. CrossRef Medline

Ruff CC, Ugazio G, Fehr E (2013) Changing social norm compliance with noninvasive brain stimulation. Science 342:482–484. CrossRef Medline

Rushworth MF, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal cortex and reward-guided learning and decision-making. Neuron 70:1054–1069. CrossRef Medline

Song XW, Dong ZY, Long XY, Li SF, Zuo XN, Zhu CZ, He Y, Yan CG, Zang YF (2011) REST: A toolkit for resting-state functional magnetic resonance imaging data processing. PLoS One 6:e25031. CrossRef Medline

Spitzer M, Fischbacher U, Herrnberger B, Grön G, Fehr E (2007) The neural signature of social norm compliance. Neuron 56:185–196. CrossRef Medline