Behavioral/Cognitive

Intention Modulates the Effect of Punishment Threat in

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

[1]Center for Brain and Cognitive Sciences and Department of Psychology, Peking University, Beijing 100871, China, [2]Singapore Institute for Neurotechnology (SINAPSE), Center for Life Sciences, National University of Singapore, Singapore 117456, Singapore, [3]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China, [4]Beijing Key Laboratory of Behavior and Mental Health, and [5]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

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

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.

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; Sneezy 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 corre-

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.

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 4 × 1 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.

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, aversio

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 ($F_{(1,23)}$

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.

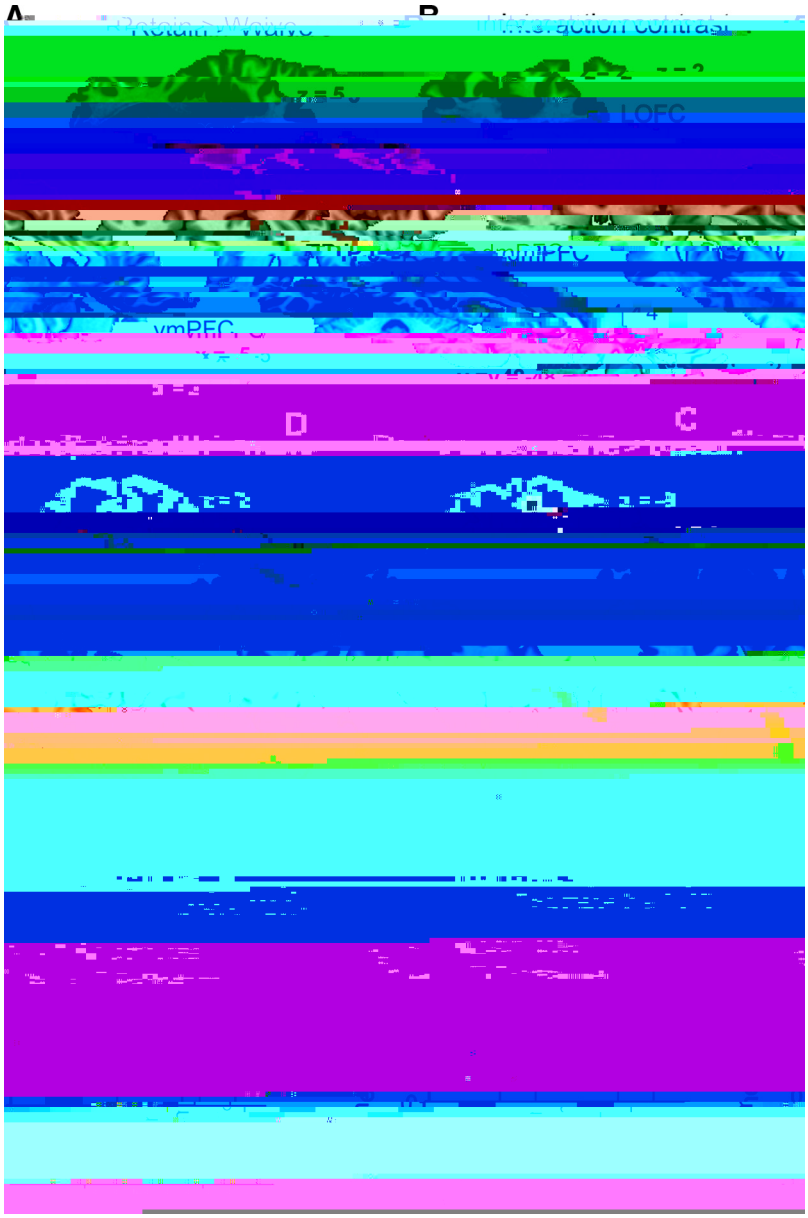

Figure 3. Analysis of brain activation. A, The whole-brain main effect contrast (Partner_Retain revealed activation in the areas typically associated with intentional/mentalizing processing (e.g., dmPFC, TPJ) and affe thalamus, dorsal caudate). B, The whole-brain interaction contrast (Partner_Waive Partner_Retain Computer_Retain) revealed activation in the bilateral LOFC and the vmPFC. C, Parameter estimates (Computer_Waive revealed activation in the bilateral LOFC and the left caudate. D, The left caudate Partner_Retain revealed activation in the rLOFC and the vmPFC. E, F, Quantitative analysis on the activation in the rLOF (Spitzer et al., 2007) and the vmPFC (Hare et al., 2009) based on the previous literature. No activation was found for P

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.

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

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 pat((905 -1.259c8 0 TD l(hodal5TD einn [(this

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

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.

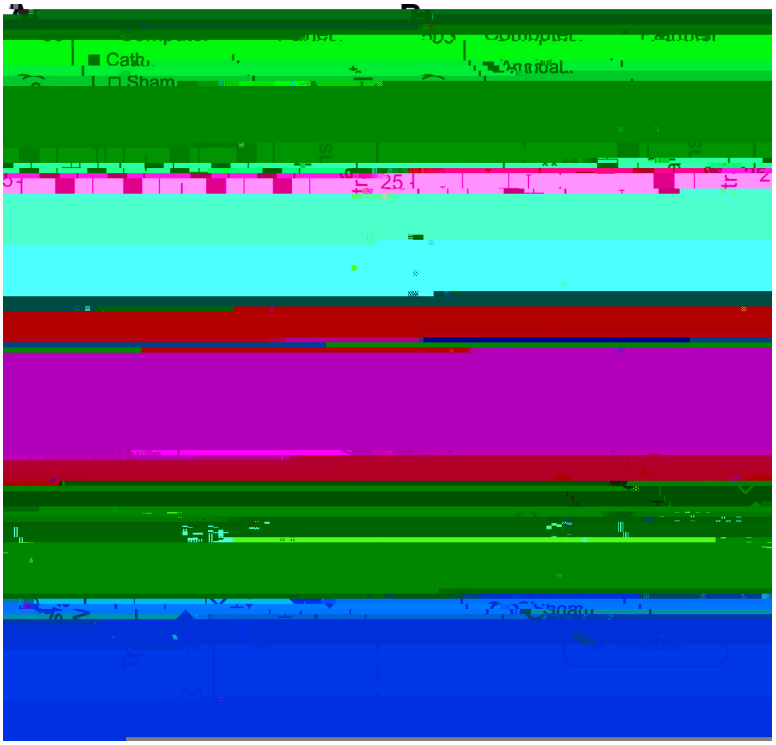

Figure 5. Results of the HD-tDCS experiments. The allocation as the function of Stimulation Type (A Sham), Decider (Computer vs Partner), and Threat (Retain vs Waive) in Experiment 1. B, Allocation in Experiment 2. C, Cathodal stimulation decreased, whereas anodal stimulation increased, the effect of punishment th transferred in the Waive condition minus the amount transferred in the Retain condition) in the int context. Error bars indicate SEM. *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