Sex-based Differences in Right Dorsolateral Prefrontal Cortex Roles in Fairness Norm Compliance

Wanting Chen\textsuperscript{a}, Shuyue Zhang\textsuperscript{b}, Ofir Turel\textsuperscript{c,d}, Youqing Peng\textsuperscript{a},

Hong Chen\textsuperscript{a*}, Qinghua He\textsuperscript{a,d,e,f,g*}

\textsuperscript{a}Faculty of Psychology, Southwest University, Chongqing, China
\textsuperscript{b} Faculty of Education, Guangxi Normal University, Guangxi Colleges and Universities Key Laboratory of Cognitive Neuroscience and Applied Psychology, Guilin, Guangxi, China
\textsuperscript{c}Information Systems and Decision Sciences, California State University, Fullerton, CA, United States
\textsuperscript{d}Brain and Creativity Institute, University of Southern California, Los Angeles, USA
\textsuperscript{e}Key Laboratory of Cognition and Personality, Ministry of Education, Southwest University, Chongqing, China
\textsuperscript{f}Southwest University Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University, Chongqing, China
\textsuperscript{g}Chongqing Collaborative Innovation Center for Brain Science, Chongqing, China

Short Title: Sex-based Differences in rDLPFC in Fairness Norm Compliance

*Correspondence should be sent to:

Prof. Hong Chen
Faculty of Psychology
Southwest University
Chongqing, 400715, China
Email: chenhg@swu.edu.cn

OR Prof. Qinghua He
Faculty of Psychology
Southwest University
Chongqing, 400715, China
Email: heqinghua@swu.edu.cn

Word Count: Abstract: 158; Introduction: 815; Discussion: 1556; Total: 4298

No. of Tables: 1; No. of Figures: 6; No. of Appendix: 4.
Abstract

Social norms are a common motivator or de-motivator of behavior through the need of humans to comply with acceptable behaviors of their social groups. Converging evidence from functional neuroimaging and noninvasive brain stimulation studies suggest that the right dorsolateral prefrontal cortex (rDLPFC) is involved in social norm compliance. Extending this view, we suggest that rDLPFC may not act uniformly when men and women face different types of fairness norm compliance situations, and that there may be a sequential update of norm compliance after experiencing punishment. Using High-Definition Transcranial Direct Current Stimulation (HD-tDCS) on 100 participants playing a sequence of economic game trials, the current study showed that (1) the right dorsolateral prefrontal cortex (rDLPFC) is involved in both voluntary and sanction-induced fairness norm compliance, (2) it has an opposite role in the two types of compliance, (3) there are sex-based differences in rDLPFC roles in fairness norm compliance, and (4) post-punishment fairness compliance is reduced compared to baseline.

Key words: fairness norm compliance; rDLPFC; HD-tDCS; gender differences; sequential updating, learning
Introduction

Humans live in a social context that frequently requires interactions and cooperation with others, including with unrelated strangers (Fehr & GaÈchter, 2002; McAuliffe, Blake, Steinbeis, & Warneken, 2017). Social norms are used in this context as a mental shortcut that can guide humans’ behaviors by setting boundaries around what is socially acceptable versus less acceptable. Social norms represent a code of conduct that is formed through reinforcement by social interactions and feedback. Such learning takes place at early stages of development (Blake & McAuliffe, 2011; Sloane, Baillargeon, & Premack, 2012), and this “mental representations of appropriate behavior” guide action throughout life (Aarts & Dijksterhuis, 2003; Liu et al., 2017). Importantly, social norms can be internalized into personal consciousness and can consequently be followed, even without external rewards or punishments (Norton & Alwang, 1993).

Social norms can take a variety of forms, ranging from moral valence to laws (J. W. Buckholtz & R Marois, 2012). Therefore, they can promote social development by influencing human social behavior at different levels. For example, at a moral level, the establishment of a fair value will make humans show voluntary fair behavior. In contrast, laws that define norms may restrain people and establish social order by punishing unfair behavior (M. Spitzer, Fischbacher, Herrnberger, Gron, & Fehr, 2007; M Spitzer, Fischbacher, Herrnberger, & Grön, 2007). Given their broad implications for many behaviors (e.g., paying taxes, donating, civilized behavior) social norm compliance has become an increasingly promising research area that can boost the building of a harmonious society (Civai & Ma, 2017).

Research has found that social norm compliance is rooted, in part, in brain networks that govern social behavior and decision making (J. W. Buckholtz & R Marois, 2012). Inputs from these networks are integrated in the prefrontal cortex (Kuehne, Heimrath, Heinze, & Zehle, 2015; Liu et al., 2017; Steinmann et al., 2014), and especially the right dorsolateral prefrontal cortex (rDLPFC). Specifically, the rDLPFC serves as a hub that incorporates inputs from various brain regions (e.g., the
thalamus, hippocampus, and dorsal striatum) and mobilizes decision responses (Hu et al., 2017; Wallis & Miller, 2003), including in situations that require cognitive flexibility (Kaplan, Gimbel, & Harris, 2016). Indeed the rDLPFC is activated in moral and social cognition tasks (Civai & Ma, 2017; Wout, Kahn, Sanfey, & Alemanc, 2005), including in social norm compliance.

The activity of the rDLPFC can be modulated through neural excitability operationalized with anodal Transcranial Direct Current Stimulation (tDCS). This is relevant for norm compliance as the increased excitability has been shown to reduce voluntary transfers (when the receiving party just accepts the transfer) and increase sanction-induced transfers (when the receiving party can "punish" the proposer for what is perceived to be an unfair transfer) in females playing an experimental game (Ruff, Ugazio, & Fehr, 2013). This game required one player to follow social norms of fair/equal division, and share a portion of her freely received tokens with a stranger. We seek to validate these results and extend them to males. Replicating these results, this time will males and females, we expect that (H1) the rDLPFC will be involved in both voluntary and sanction-induced fairness norm compliance in this experimental game, and (H2) it will have opposite roles in the two types of compliance (voluntary vs. sanction-induced).

We next postulate that males and females could differ in their neural and behavioral responses to stressful situations (Lighthall, Mather, & Gorlick, 2009; Lighthall et al., 2012; Mather & Lighthall, 2012), which can result in differences in prosocial and/or aggressive behaviors (Chen & Abedin, 2014; Richardson, Hammock, Smith, Gardner, & Signo, 1994). Specifically, given sex-based differences in altruistic prosocial preferences (Andreoni & Vesterlund, 2001), risk-taking behaviors (Donovan et al., 2018) and utilitarian judgment (Fumagalli et al., 2010), it is reasonable to expect that the abovementioned patterns, which have been observed in females, may change and even possibly reverse in males. For example, females show reversed patterns of activation compared to males in social dilemma and trust games, in regions that govern risky and social decision making (Rodrigo, Padrón, De Vega, & Ferstl, 2014) including the prefrontal cortex (Garbarini et al., 2014). In addition, it has been
shown that women encode sharing money more strongly than men do (Soutschek et al., 2017), and men show greater retaliatory aggressions when a threat of social or monetary punishment is present (Chen & Abedin, 2014; Richardson et al., 1994). Such differences are reflected in brain activity, including in prefrontal regions (Ding et al., 2017). Women tend to engage the rDLPFC more than men do, in response to rewards they need to inhibit (Cornier, Salzberg, Endly, Bessesen, & Tregellas, 2010). Moreover, empathy to other players, which tends to be stronger in women (Eisenberg & Lennon, 1983), alters the effect of noninvasive brain stimulation on the rDLPFC (Brüne et al., 2012). We hence expect that (H3) there will be sex-based differences in rDLPFC roles in fairness norm compliance. These may even manifest in opposing effects of tDCS on the rDLPFC; while in women anodal stimulations is expected to reduce voluntary transfers and increase sanction-induced transfers (Ruff et al., 2013), in men, who are presumed to be less empathic, more utilitarian focused, and less altruistic and prosocial (Soutschek et al., 2017), rDLPFC stimulation may result in different outcomes, possibly even increased voluntary transfers and decreased sanction-induced transfers.

In addition, we seek to extend the abovementioned research by considering the learning effect of sanctions/punishments. Learning theories suggest that rational people incorporate prior experiences into current cognitions, which then inform their behavioral choices (Heininga, van Roekel, Wichers, & Oldehinkel, 2017). For example, children who are punished for incorrect answers via a loud tone learn faster than others to correct their behavior (R. K. Penney & A. A. Lupton, 1961). In our case, punishing for the division of the tokens in prior rounds is likely to signal to participants that their allocations were not socially acceptable (too low). This should motivate them to ramp-up the number of tokens they share with their game partner in subsequent rounds. We therefore expect that (H4) token transfers in post-punishment rounds will be higher compared to token transfers in the baseline (pre-punishment) rounds.
Materials and methods

Participants

One hundred and one right-handed Chinese college students volunteered to participate in the study. One of them was excluded from further analysis because he transferred the same amount of money to his partner in all trials, leaving 100 participants (50 females and 50 males, 18-25 years old, mean age = 20.5 years old). All participants had normal or corrected-to-normal vision, and none of them reported present or past neurological or psychiatric conditions. All participants gave written informed consent to the experiment procedures, which were approved by the Institutional Review Board of Southwest University. All participants were naïve to the tDCS procedure. The overall compensation depended on the tokens they earned in the economic game, but the minimum payment for each participant was set to 30 yuan.

Experimental paradigm and Measures

Preliminary Survey. Before the sequence of economic games, participants completed an online questionnaire capturing Machiavellism (Chinese version of Mach-IV scale), empathy (Chinese version of Interpersonal Reactivity Index-C), anxiety (Chinese version of the STAI scale) and risk-taking attitudes (the Chinese version of DOSPERT scale). The translated instruments have been shown to be valid and reliable in Chinese speaking populations (Li & Qian, 1995; Tang & Hao, 2011; Wu & Cheung, 2014; Zhang, Dong, & Wang, 2010). Due to randomization in condition assignment, we expected that such control variables do not differ between the anodal and sham groups.

Experiment. We conducted a single-blind, placebo-controlled, between-subject design, following a procedure similar to this used in Ruff et al. (2013). This paradigm is well suited for the study of fair behaviors, because it balances between experimental control and psychological processes (M. Spitzer et al., 2007; Vavra, van Baar, & Sanfey, 2017; H. Zheng & Zhu, 2013). It builds on the notion that people acquire fairness norms that can be activated with no incentives or punishments; and
these norms normally see a relatively equal split (equality principle of justice) or a split that reflects relative effort (equity principle of justice) as fair (Johnston, 2011; Rawls, 2009). This parading is expected to activate the DLPFC given its role in self-control. In the voluntary behavior condition, self-control likely reflects the inhibition of prepotent alluring responses (e.g., to give nothing). Adding sanctions complicates this decision-process, because participants may consider the intention behind the sanction and expected opponent behaviors. Thus, the sanction condition elicits social norm and opponent behavior considerations, both of which are expected to engage the rDLPFC.

Participants were randomly assigned to one of two stimulation conditions: Anodal (25 females and 25 males) or sham (25 females and 25 males) stimulation of the rDLPFC. Participants engaged via computer terminals in an economic game that had real financial consequences. In this game (Figure 1), participants took the role of Player A repeatedly and were randomly paired with a different player B in each round. Player B was always a preprogrammed computer. Nevertheless, participants were informed that they are facing a human player (another classmate). We tested the verisimilitude of this procedure in a pilot study with 56 volunteers (different from those recruited for the main study). It revealed that 91.1% of participants believed that at least one of the opponents is a real person. Hence, the decision-making context in the present experiment was somewhat social and as such, deviates, in part, from the context in Ruff et al. (2013). See Supplementary Information, Appendix A, for details.

In every round, both players had an initial endowment of 25 tokens. In addition, participants received another 100 of tokens and proposed how many tokens to transfer to Player B. The proposed transfer could be any integer product of 10, from 0 to 100 (e.g., 0, 10, and 70). Each player participated in two variations of the game organized in three blocks. In baseline rounds (pre-punishment) and relative baseline rounds (post-punishment) (Figure 1.A), participants (Player A) proposed a transfer and player B was forced to accept it. In punishment/sanction rounds (Figure 1.B), participants were informed that Player B has the option to refuse the proposal and use the initial endowment of 25 tokens to punish participants by reducing proposers’ tokens. Each
participant performed three blocks (Figure 2.B), including baseline rounds, punishment rounds and relative baseline rounds (same as baseline rounds, just post-punishment). Each block consisted of 12 rounds of the same type, with the prevailing trial type and rules indicated at the beginning of each round (Figure 3).
It was assumed that fairness norms prescribe a close to equal split of the “cake” of tokens as fair (Rawls, 2009). This notion was used for programming Player B, the probability of it refusing the offers and how many tokens it invests into punishment. These preset probabilities are given in Supplementary Information, Appendix B. Specifically, for every token that Player B chose to invest in punishing Player A, Player A’s earnings were reduced by 5 tokens. To illustrate, if Player A transfers nothing to B, and consequently has 25+100=125 tokens, and B has 25 tokens, then B can invest 25 tokens into punishment and reduce player A’s earning to zero.

Submitting a proposed token transfer was operationalized by typing the number of tokens and pressing the "Enter" key on the computer keyboard. In the first block, token transfers can be interpreted as voluntary transfers to Player B, which are based on pure social norm compliance. In the second block, the punishment threat may induce A to give more in order to prevent being punished by player B. As a
consequence, the difference in token transfer amounts between punishment and baseline rounds can be considered Player A's sanction-induced norm compliance as it captures the added value of sanctions beyond baseline norm compliance. In the third block, the punishment information from the previous rounds informs participants; hence, token transfers in this block relative to the initial block (baseline rounds) can demonstrate how the voluntary compliance of participants has changed after punishment.

Stimulation Protocol

During the experiment, we applied high-definition transcranial direct current stimulation (HD-tDCS), which provides more targeted current than tDCS (Gözenman & Berryhill, 2016) to participants' rDLPFC. HD-tDCS was delivered by connecting a 4 × 1 multichannel stimulation adapter and a battery-driven stimulator (Soterix Medical Inc., New York, NY, USA) and by following recommended HD-tDCS procedures (Villamar et al., 2013). The tDCS temporarily alters cortical excitability and activity in target neurons (Woods et al., 2016; Zhao et al., 2017). We expected that the anodal electrode would enhance cortical excitability (Nitsche & Paulus, 2000) as demonstrated in prior applications of tDCS to the rDLPFC (Ruff et al., 2013).

We specifically applied anodal or sham tDCS over the rDLPFC region, which is equal to the right F4 according to the International EEG 10-20 system. Four cathodal electrodes sat equidistantly around the anode (F2, AF4, F6, FC4, see Supplementary Information for details. Appendix C.). Anodal active stimulation was 1.5mA. Active brain stimulation has been continued until the participants had finished the economic game (less than 30 min in all cases). Because the impact of tDCS on brain excitability stabilizes after several minutes (Nitsche & Paulus, 2000), tDCS was applied for three minutes without any task, before participants started the economic game blocks. Sham stimulation was applied using the same procedures, but stimulation lasted only for the first 30s to create a potential itch equivalent to what anodal tDCS subjects feel. After this initial ineffective current stimulation, the current stimulator was de-ramped. Such sham stimulation has been shown to be reliable (Gandiga, Hummel, & Cohen, 2006).
Indeed, the effect of the stimulation was indistinguishable to the participants. To demonstrate this, immediately after the experiment, participants completed a questionnaire in which they expressed how much they perceived the stimulation to affect their behavior (ranging from 1 “not at all” to 7 “extremely”). We also measured participant tolerance of brain stimulation (ranging from 1 “totally untolerance” to 7 “extremely acceptable”). In addition, before and after the behavioral task, all participants completed the Multidimensional Mood State Questionnaire (MMSQ, including mood, alertness, and calmness) to make sure the different tDCS manipulations did not affect emotional states (Figure 2). Lastly, we wanted to ensure that the tDCS did not affect the participants’ fairness perception. To this end, we measured fairness beliefs, anger expectations related to how Player B is expected to feel after their offers, and how many punishment tokens Player B should invest at different transfer levels.

**Statistical analysis**

All statistical analyses were performed with SPSS 24.0. To assess the effects of tDCS on voluntary fairness compliance and sanction-induced fairness compliance, as well as differences between females and males, we used univariate analysis and entered personality variables as covariates into the analysis. Gender and brain stimulation type were entered as fixed factors into the analysis, and round type was included as a within-subject factor. Participants’ pre- and post-stimulation emotions were analyzed with paired-samples t-tests. Independent samples t-test was used to compare the fairness beliefs of participants exposed to anodal vs. sham brain stimulation. Partial Eta-squares were computed to measure effect sizes.

**Results**

**Experiment**

Descriptive statistics are given in Table 1. Results are depicted in Figures 4, 5, and 6. Average transfers by sex, stimulation condition and game block are given in
Table 1. ANCOVA for voluntary fairness compliance as the outcome exhibited a significant interaction \((F(1,91) = 4.356, p = .040; \eta^2_p = .046)\) between sex and stimulation type (Figure 4.A). Post hoc contrast analyses revealed that this difference, between sham group and anodal, was approaching significance in both females \((F(1,43) = 3.241, p = .079; \eta^2_p = .070)\) and males \((F(1,43) = 3.546, p = .066; \eta^2_p = .076)\). Similarly, ANCOVA for sanction-induced fairness compliance as the outcome has demonstrated that the interaction between sex and stimulation type was statistically significant \((F(1,91) = 4.907, p = .029; \eta^2_p = .051)\) (Figure 4.B). Post hoc contrast analyses indicated that this interaction was driven by the anodal stimulation. There was approaching significance main effect of sex in anodal stimulation \((F(1,43) = 3.979, p = .052; \eta^2_p = .085)\). It enhanced sanction-induced fairness compliance in women \((F(1,43) = 4.488, p = .040; \eta^2_p = .095)\), but not in men \((F(1,43) = 2.160, p = .149; \eta^2_p = .048)\). However, there were no significant main or interaction effects (sex and stimulation type: all \(F(1,91) < 1.826, all p > .180\)) on voluntary fairness compliance after punishment. ANCOVA for the difference of token transfers in the relative baseline (post-punishment) rounds and in the baseline (pre-punishment) rounds demonstrated no significant difference \((F(1,91) < 1.522, all p > .220)\) (Figure 4.C).
Control Variables

Summary of statistical testing for possible stimulation group differences in
control variables are given in Supplementary Materials, Appendix D. There was no significant difference between pre- and post-stimulation emotional states in the anodal group (good-bad mood \( t = .489, p = .627 \) and calmness-restlessness \( t = .155, p = .877 \)). Furthermore, no significant main or interaction effects of stimulation group and sex on emotional states were observed (all \( F(1,91) < 1.431, all \ p > .235 \)). All participants were clearly aware of the notion that higher (close to 50%) transfers are fairer, leading to smaller expected punishment by and less anger in Player B (Figure 5). Importantly, no significant differences in participants’ beliefs between the anodal and the sham groups (all \( F(1,98) < 1.99, all \ p > .161 \), and between males and females (all \( F(1,98) < 2.228, all \ p > .139 \)) were detected. The degree to which participants perceived the stimulation to affect their behavior was low (total mean: 2.13 [1 to 7]; almost all [87.1%] reported a score below 4). Similarly, participants from the sham and anodal stimulation groups gave statistically indistinguishable ratings of tDCS’s acceptance (total mean: 6.67 [4 to 7], almost all [76.2%] reported scores of at least 6). Most notably, there were also no significant differences in tDCS’s acceptance between males and females (all \( F(1,98) < .156, all \ p > .693 \)). Taken together, these potential confounds of tDCS were not influential in this study and cannot explain the difference in participant behaviors.
Discussion

Fairness is a complex concept with major societal, business, political and individual implications. As such, it has been extensively studied in multiple disciplines, such as psychology, sociology, and economics (Adams, 1965; Leventhal, 1980). One key pillar of fairness perception is equality or the social norm that people deserve a similar share of the pie if no effort is involved (Xiao & Bicchieri, 2010). This principle acts as an objective benchmark in investigations of fairness (McAuliffe et al., 2017). Consequently, people develop a general social norm that a 50-50 split of freely received money is fair (Ruff et al., 2013). In the present experiment, we employed an economic game in which fairness norms are expected to be followed (voluntarily or under punishment threats), and investigated how HD-tDCS applied to the rDLPFC can alter fairness norm compliance. The objectives were to replicate the anodal stimulation effects results in Ruff et al. (2013) and to extend them in two
directions. First, we hypothesized and tested between-sex differences in rDLPFC stimulation effects on fairness norm compliance. Second, we hypothesized and examined a learning effect by implementing a post-punishment voluntary compliance phase. These are important extensions of Ruff et al. (2013) because they present a more realistic and complete picture: many economic decisions are made by man and/or women, and many decisions are made post punishment (e.g., paying taxes after a penalty was levied in the previous year).

The results of our study provide interesting insights. First, we generated a chart that captures transferred in all conditions in every round (Figure 6). We found that participants transferred less in the voluntary condition (Baseline rounds). The transfers increased when a sanctioning threat became present (Punishment condition), in both men and women. There does not seem to be a learning pattern in each condition. This may be seen as reflecting the pre-existence of fairness norms, which support our assumption that people act based on innate and socially acquired fairness norms, even without sanctions and without on-the-spot learning.

Second, they confirm the prior findings in Ruff et al. (2013) and partly support H1 and H2. We found that the rDLPFC is involved in voluntary fairness norm
compliance and that its excitability plays opposite roles in voluntary vs. sanction-induced compliance. Compared with females, the excitability of rDLPFC impact on males in both types compliance, but have not statistically significantly changed in sanction-induced compliance. These results are consistent with Ruff et al. (2013), but also extend the generalizability of prior research to males, and possibly from European to Chinese populations.

Third, when the excitability of rDLPFC was enhanced in women by anodal tDCS stimulation, the level of their voluntary fairness norm compliance has decreased compared to the sham stimulation group. In contrast, their sanction-induced transfer has increased compared to the sham stimulation group. This finding is consistent with Ruff et al. (2013). Extending this view, we showed that the effects of anodal tDCS on different types of compliance were also opposite in direction among men, but were reversed compared to the effects observed in women. Specifically, the voluntary transfer in men has increased under tDCS compared to sham, while the sanction-induced transfer has decreased under tDCS compared to sham (Figure 4). This highlights possible sex-based differences in rDLPFC roles in voluntary vs. sanction-based fairness norm compliance and supports H3. This reaffirms the idea that the sexes can differ in the way they cognitively and in terms of neural activation assess and act in social situations (Garbarini et al., 2014), including in monetary decisions (Soutschek et al., 2017).

While all regions of the PFC can be involved in decision processes (Asgher et al., 2017), the DLPFC integrates information from other regions and selects context-appropriate behavioral responses during norm-based decision-making processes (J. W. Buckholtz & R. Marois, 2012; Buckholtz et al., 2015; Hu et al., 2017). Hence, the observed sex-based differences in how tDCS over rDLPFC modulates fairness norm compliance may be informative and can be used for better understating decision making in men and women. Given that the striatum shows enhanced activation for prosocial rewards in females and for selfish rewards in males (Soutschek et al., 2017), and the rDLPFC integrates information from the striatum into such decisions (Kohno, Morales, Ghahremani, Hellemann, & London, 2014; Leh,
Ptito, Chakravarty, & Strafella, 2007; McClure, Ericson, Laibson, Loewenstein, & Cohen, 2007; Wallis & Miller, 2003; Yamamoto, Woo, Wager, Regner, & Tanabe, 2015; Yoder & Decety, 2014), one possibility is that the rDLPFC suppresses these dominant preferences and thus leads to opposite behaviors in females and males if transfers are voluntary. However, the existence of sanctions complicates decision-making and integrates fairness norms with motivation to reduce the possibility of being punished (Gianotti, Nash, Baumgartner, Dahinden, & Knoch, 2018; L. Zheng et al., 2017). Strategic behavior has been positively associated with the activation in DLPFC because the DLPFC mediates suppression of prepotent responses as a means to achieve set goals (Gianotti et al., 2018; Strang et al., 2014). That is, the more complex set of goals (complying with norms plus avoiding penalties) can explain stronger engagement of the rDLPFC. Compared to voluntary condition, the DLPFC may be more focused on balancing emotion and cognition in the punishment threat condition. This explanation, though, cannot be supported with our data and requires further research. We therefore call for future research to examine such explanations; delving into sex-based differences in rDLPFC activation is important as findings can be used as a basis for developing sex-specific interventions in cases in which the rDLPFC is impaired (Pa et al., 2014).

Fourth, in contrast to our expectations for an increase, compared with baseline transfers, transfers slightly (but not significantly) decreased after punishment (Figure 4.C). One possible explanation for this is that post punishment decisions are influenced not only by cognitive updates (Ding et al., 2017; Heininga et al., 2017; R. K Penney & A. A Lupton, 1961) but also by post-punishment emotions (Bechara, 2004). Losses, as in the case of punishment, can induce negative emotions, in which case the proposer may give more stingy transfers in the next round (Yang & Hao, 2014). Stringer transfers may also stem from the need to retaliate against the punishment imposed by Player B (Dreber & Rand, 2012; Sigmund, 2007), which can be enacted as a means to restore emotional homeostasis, maintain self-image and resolve cognitive dissonance. Therefore, H4 was not supported. It implies that continued fair behavior cannot only be achieved by a one-time punishments.
Fifth, as per the behavioral data of the sham group (Table 1), the voluntary transfer amounts of women were higher than those of men. Although these differences were not statistically significant, the observed trend is in line with prior findings indicating that women tend to show more equitable and prosocial attitudes and behaviors compared to men (Heilman & Chen, 2005). Such differences may stem from cultural expectations and gender stereotypes (Eagly & Crowley, 1986; Soutschek et al., 2017). In line with fairness principles when threats are present (Ruff et al., 2013; Manfred Spitzer, Fischbacher, Herrnberger, Grön, & Fehr, 2007), we found that when a sanctioning threat is present, the transfer increases compared to baseline, in both men and women. This means that men and women are sensitive to the punishment threat and show more normative conforming behavior when sanctions are feasible.

Sixth, it is interesting to consider differences in transfers between this study and those observed in the social context of Ruff et al. (2013). Because Ruff et al. (2013) included only women, we compare the transfers made by women. In Ruff et al. (2013) the voluntary and sanction-induced transfers were about 15 and 29, correspondingly, in the sham (no) stimulation condition; and 10 and 38, correspondingly, in the anodal stimulation condition. In our study they were 23.43 (95% CI [16.93, 29.82]) and 37.03 (95% CI [33.30, 40.80]), correspondingly, in the sham (no) stimulation condition; and 16.67 (95% CI [11.56, 21.84]) and 38.37 (95% CI [35.83, 40.99]), correspondingly, in the anodal stimulation condition. Thus, except for the sanction-induced transfers under anodal stimulation, the transfers in our study seem larger. This may reflect cultural norm differences between Western and Chinese populations (Han, Glover, & Jeong, 2014). Chinese participants may arguably be more collectivistic compared to largely individualistic Westerners (Hofstede, Hofstede, & Minkov, 2010). This can also be explained by the perceived realism of the games and anthropometrism of player B, which may have differed between the studies. The fact that both samples reached a similar level of transfers after anodal stimulation may reflect a ceiling effect representing a maximal rational amount people are willing to share. These ideas, though, merit further research.
Although this study provides important initial insights, there are also limitations that should be acknowledged. First, the spatial resolution of tDCS to manipulate neural activity is limited; tDCS may also activate neurons in adjacent regions (Asgher et al., 2017; Bogdanov, Ruff, & Schwabe, 2017; Haller & Schwabe, 2014). While this is mitigated by the use of HD-tDCS, we still air on the cautious side and conclude that the rDLPFC is necessary for fairness norm compliance but that norm compliance is not restricted to neural activity within this area. Second, participants were confronted with a preprogrammed computer that was presented as another human. Although we created an impression of a social context in which the interaction is with another human, we cannot refute the possibility that the situation was seemed as nonsocial for some participants. When confronted with a virtual human, participants in prior research were sensitive to punishment threats, but the effects of tDCS on sanction-induced transfers were weaker than those during interactions with a human player (Ruff et al., 2013). Hence, future research may put stronger emphasis on the decision making context and extend our finding to different social contexts and levels of anthropometrism.

To conclude, the present findings confirm that the DLPFC is a key brain region involved in fairness norm compliance; but also show that it plays different roles in men and women, and in types of norm compliance (voluntary vs. sanctioned), in such decisions. We specifically show that anodal stimulation over the right DLPFC enhances voluntary fairness norm compliance in males while decreases it in females. In contrast, activating the right DLPFC decreases sanction-induced fairness norm compliance in males while enhances it in females. Moreover, slight and not significant decrease in voluntary fairness norm compliance was observed after experiencing punishment. This contradicts learning theories (Rescorla & Solomon, 1967; Waelti, Dickinson, & Schultz, 2001), and suggest potential sequential emotional retaliation or image-restoration effect, which requires further research.
Acknowledgments:

QH was supported by research grants from the National Natural Science Foundation of China (31400959), Entrepreneurship and Innovation Program for Chongqing Overseas Returned Scholars (cx2017049).

Conflict of Interest: The authors declare no competing financial interests.

Reference


Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one’s political beliefs in the face of counterevidence. Scientific Reports, 6, 39589. doi:10.1038/srep39589

https://www.nature.com/articles/srep39589#supplementary-information


**Figure Captions**

**Fig.1:** Economic game used to measure fairness norm compliance.

In each round, both players have an initial endowment of 25 tokens and receive additional 100 tokens. Player A is always the proposer and sends a transfer of X tokens to Player B from the 100 tokens. At the end of the experiment, we randomly selected a round from the 36 rounds and 10% of the participants’ proceeds of this round were converted into real money. In other words, the earnings of the participants depended on the tokens they retained in the economic game.

(A) **Baseline round:** Transfer of X tokens is implemented as proposed; it represents voluntary fairness norm compliance. (B) **Punishment round:** Whether the proposal passed or not is based on random feedback from Player B. If Player B accepts, the transfer is implemented. If not, Player B can refuse and invest Y tokens from the initial endowment to punish Player A by reducing his/her payoff by 5*Y tokens. The round type was indicated at the beginning of each trial. Hence, in punishment trials Player A is expected to realize the possible sanctions. The addition in transfer in
punishment trials compared to the baseline rounds reflects the added sanction effect (i.e., sanction-induced fairness norm compliance).
Fig.2. The experimental process

(A) The experimental process in the laboratory: First, participants were informed regarding the experimental process and task. They were then asked to complete the Multidimensional Mood State Questionnaire and several personalities and demographics scales. Then, they performed three blocks including 36 trials. Immediately after the experiment, they were asked to complete a series of questionnaires including the Multidimensional Mood State Questionnaire and items capturing their perceptions and beliefs. (B) The main task: Participants completed 36 rounds of game in sequence. In all rounds, they were paired with Player B, which was a computer number randomly assigned in each round. The difference between the baseline and the relative baseline round is the relative order of punishment round; the first is pre-punishment rounds and the second is post-punishment rounds.
Fig.3: Game process.

**Round type:** At the beginning of each round the participants were informed about the type of upcoming trial. Hence, participants always knew whether they faced a punishment threat or not.

**Connect to Player B:** After round type information, the screen showed “waiting, connecting to Player B” in order to increase the authenticity. To further increase authenticity, we randomly set five connection failures in all trials. The fictitious connection waiting time was a random number [4800ms - 7500ms].

**Player B number:** When the connection was successful, a blank screen with the player B’s number [random: 1 to 36] was presented; each number appeared only once in all trials.

**Distribution:** The screen asked participants to enter the number of tokens to be given to player B.

**Feedback of Player B:** Accept or refuse and punish.

**Payoff:** Participants’ and Player B’s payoff were presented on the screen.

**The end:** End of the round; an invitation to start the next trial is presented.
Fig.4: rDLPFC stimulation changes voluntary and sanction-induced norm compliance of females and males.

(A) Voluntary fairness norm compliance: Average transfers for baseline rounds. (B) Sanction-induced fairness norm compliance: Average transfer difference between punishment rounds and baseline rounds. This value represents the add compliance caused by the threat (i.e., excluding voluntary compliance). (C) Voluntary fairness after punishment and voluntary fairness compliance: Average transfers for relative baseline rounds and baseline rounds in both sex. Compared with the baseline rounds, the difference represents the change in fairness norm compliance after punishment. Error bars reflect standard errors.
Fig. 5: The brain stimulation does not affect participants’ perception and beliefs. 

(A) Average rating of perceived fairness for different transfer levels (scale from 1 “very unfair” to 7 “very fair”). (B) Average rating of anticipated anger felt by player B for different transfer levels (scale from 1 “not angry at all” to 7 “very angry”). (C) Average expected payoff reduction resulting from B’s punishment.
Fig.6: The transfers in all conditions in every round.

B1-B12: 12 rounds in sequence in baseline conditions (Block 1).

P1-P12: 12 rounds in sequence in punishment conditions (Block 2).

R1-R12: 12 rounds in sequence in relative baseline conditions (Block 3).
Supplementary Information

Appendix A: The Verisimilitude of pre-programmed player B

Fifty-six participants (42 females and 14 males, 17-21 years old, mean 19.4 years) volunteered to participate in an experiment that assessed the verisimilitude of pre-programmed player B. The procedures and block game were identical to those employed in the main study, but without tDCS.

After finishing the sequence 36 trials of the economic games described in the main text, the participants assessed the experimental procedure with a questionnaire. This questionnaire captured the perceived reliability of scene setting, rigor of guidance, authenticity of the process and perceptions of whether the opponents are real people. For each question, participants were shown a sliding scale with proportional value on each side to represent the extreme values (Figure S1.). Mean values and standard deviations are summarized in Table S1. The findings indicated that participants perceived the process and setting to be quite, but imperfectly, realistic. They had mixed perceptions regarding whether the opponent is a real person. The skepticism, nevertheless, may have been an artifact that was instigated by raising this question. Overall, we concluded that the procedure is reasonably, but imperfectly realistic, like common experimental procedures.

![Figure S1](image-url)
Table S1. Descriptive statistics of experimental procedure attributes

<table>
<thead>
<tr>
<th></th>
<th>Min(%)</th>
<th>Max(%)</th>
<th>Mean(%) ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>The reliability of scene setting</td>
<td>22</td>
<td>100</td>
<td>71.27 ± 19.05</td>
</tr>
<tr>
<td>The rigor of guidance</td>
<td>10</td>
<td>100</td>
<td>62.66 ± 22.09</td>
</tr>
<tr>
<td>The authenticity of the process</td>
<td>20</td>
<td>100</td>
<td>64.46 ± 20.36</td>
</tr>
<tr>
<td>Are all the opponents real person?</td>
<td>0</td>
<td>100</td>
<td>47.16 ± 32.85</td>
</tr>
</tbody>
</table>
Appendix B: The probability of Player B refusing the offers as a function of Player A’s transfer and Player B’s tokens devoted to punishment

<table>
<thead>
<tr>
<th>Punishment</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0/100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10%</td>
<td>10%</td>
<td>80%</td>
</tr>
<tr>
<td>10/100</td>
<td>10%</td>
<td>5%</td>
<td>15%</td>
<td>20%</td>
<td>50%</td>
<td>0</td>
</tr>
<tr>
<td>20/100</td>
<td>30%</td>
<td>10%</td>
<td>20%</td>
<td>30%</td>
<td>10%</td>
<td>0</td>
</tr>
<tr>
<td>30/100</td>
<td>50%</td>
<td>30%</td>
<td>10%</td>
<td>10%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40/100</td>
<td>80%</td>
<td>20%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50/100</td>
<td>100%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Setting the probabilities and punishment amounts of Player B followed these two principles:

1. Participants were threatened that Player B has the option to refuse the proposal and use his or her initial endowment of 25 tokens to punish participants by reducing Player A’s tokens.

2. People were expected to develop a general social norm that a 50-50 split of freely received money is fair. However, not all responders would reject distribution below 50 tokens, and previous studies have indicated that about 50% of responders punish proposers who violate fairness norms by offering less than 20%--30% of the sum. To generate a credible punishment threat in the second
block, the number of tokens that could be spent on punishment was calibrated in the following pattern: when the transfer reaches 50% or more (which is considered a fair division), the punishment threat was removed; when the transfer was lower than 50%, the punishment and number of tokens that could be spent on punishment were proportional to the distance from 50%.

In summary, when participants proposed unfair offers, they may be punished. This threat of loss, was expected to alert participants to fairness violations and elicit responses.
Appendix C: Localization of stimulation brain area

We specifically applied anodal or sham tDCS over the rDLPFC region, which is equal to the right F4 according to the International EEG 10-20 system. Four cathodal electrodes sat equidistantly around the anode (F2, AF4, F6, FC4). Using 3-dimensional (3D) Patriot Digitizer, we confirmed localization of the stimulated brain areas (see Table S2).
### Table S2 Localization of stimulation brain area

<table>
<thead>
<tr>
<th>MIN</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>BA (brain area, the probability)</th>
<th>AAL (brain area, the probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC4</td>
<td>51</td>
<td>27</td>
<td>40</td>
<td>(1) 9 - Dorsolateral prefrontal cortex, 0.2179</td>
<td>(1) Right Middle Frontal Gyrus, 0.77043</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2) 44 - pars opercularis, part of Broca's area, 0.51362</td>
<td>(2) Right Inferior Frontal Gyrus, opercular part, 0.18288</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(3) 45 - pars triangularis Broca's area, 0.2607</td>
<td>(3) Right Inferior Frontal Gyrus, triangular part, 0.046693</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(4) 46 - Dorsolateral prefrontal cortex, 0.0077821</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>30</td>
<td>50</td>
<td>41</td>
<td>(1) 9 - Dorsolateral prefrontal cortex, 0.79186</td>
<td>(1) Right Superior Frontal Gyrus, dorsolateral, 0.44344</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2) 46 - Dorsolateral prefrontal cortex, 0.20814</td>
<td>(2) Right Middle Frontal Gyrus, 0.55656</td>
</tr>
<tr>
<td>F4</td>
<td>45</td>
<td>45</td>
<td>32</td>
<td>(1) 9 - Dorsolateral prefrontal cortex, 0.036364</td>
<td>(1) Right Middle Frontal Gyrus, 0.98636</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2) 45 - pars triangularis Broca's area, 0.45455</td>
<td>(2) Right Inferior Frontal Gyrus, triangular part, 0.013636</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(3) 46 - Dorsolateral prefrontal cortex, 0.50909</td>
<td></td>
</tr>
<tr>
<td>F6</td>
<td>55</td>
<td>39</td>
<td>19</td>
<td>(1) 45 - pars triangularis Broca's area, 0.9639</td>
<td>(1) Right Middle Frontal Gyrus, 0.31047</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2) 46 - Dorsolateral prefrontal cortex, 0.036101</td>
<td>(2) Right Inferior Frontal Gyrus, triangular part, 0.68953</td>
</tr>
<tr>
<td>AF4</td>
<td>40</td>
<td>58</td>
<td>20</td>
<td>(1) 10 - Frontopolar area, 0.20084</td>
<td>(1) Right Superior Frontal Gyrus, dorsolateral, 0.11297</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2) 45 - pars triangularis Broca's area, 0.0083682</td>
<td>(2) Right Middle Frontal Gyrus, 0.88703</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(3) 46 - Dorsolateral prefrontal cortex, 0.79079</td>
<td></td>
</tr>
</tbody>
</table>
Appendix D: Summary of statistical testing of possible stimulation group differences in control variables (Table S3, Table S4).

**Table S3. Summary of statistics testing for personality variables.**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Factor</th>
<th>Machiavellism</th>
<th></th>
<th>Empathy</th>
<th></th>
<th>Anxiety</th>
<th></th>
<th>Risk-taking attitudes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F value</td>
<td>p value</td>
<td>F value</td>
<td>p value</td>
<td>F value</td>
<td>p value</td>
<td>F value</td>
</tr>
<tr>
<td>Female</td>
<td>tDCS</td>
<td>.380</td>
<td>.541</td>
<td>.280</td>
<td>.599</td>
<td>.023</td>
<td>.879</td>
<td>1.147</td>
</tr>
<tr>
<td>Male</td>
<td>tDCS</td>
<td>2.112</td>
<td>.153</td>
<td>.045</td>
<td>.833</td>
<td>1.979</td>
<td>.166</td>
<td>.390</td>
</tr>
<tr>
<td></td>
<td>tDCS</td>
<td>2.391</td>
<td>.125</td>
<td>.273</td>
<td>.603</td>
<td>1.192</td>
<td>.278</td>
<td>1.444</td>
</tr>
<tr>
<td>Combined</td>
<td>Gender</td>
<td>3.030</td>
<td>.082</td>
<td>1.209</td>
<td>.274</td>
<td>.396</td>
<td>.531</td>
<td>3.099</td>
</tr>
<tr>
<td></td>
<td>tDCS*gender</td>
<td>.716</td>
<td>.400</td>
<td>.048</td>
<td>.826</td>
<td>.763</td>
<td>.385</td>
<td>.107</td>
</tr>
</tbody>
</table>

Table S3. Summary of statistics testing for personality variables. We compared scores on several personality scales for participants between females and males. We conducted ANOVAs with the independent variables tDCS (anodal vs. sham) and gender (female vs. male). The groups did not significantly differ in any of these variables.
Table S4. Summary of statistics testing for general emotional state

<table>
<thead>
<tr>
<th></th>
<th>Good-Bad Mood</th>
<th>Alertness-Tiredness</th>
<th>Calmness-Restlessness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t value</td>
<td>p value</td>
<td>t value</td>
</tr>
<tr>
<td>tDCS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sham</td>
<td>1.283</td>
<td>.206</td>
<td>1.505</td>
</tr>
<tr>
<td>Anodal</td>
<td>.489</td>
<td>.627</td>
<td>1.845</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1.139</td>
<td>.260</td>
<td>.809</td>
</tr>
<tr>
<td>Male</td>
<td>.777</td>
<td>.441</td>
<td>2.358</td>
</tr>
<tr>
<td>Combined</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sham*female</td>
<td>1.028</td>
<td>.314</td>
<td>.296</td>
</tr>
<tr>
<td>Sham*male</td>
<td>.877</td>
<td>.389</td>
<td>1.569</td>
</tr>
<tr>
<td>Anodal*female</td>
<td>.610</td>
<td>.548</td>
<td>.786</td>
</tr>
<tr>
<td>Anodal*male</td>
<td>.041</td>
<td>.968</td>
<td>1.943</td>
</tr>
</tbody>
</table>

Table S4. Summary of statistics testing for general emotional state. Participants’ pre- and post- stimulation emotions were measured by three subscales indexing mood, alertness, and calmness. We computed paired-samples t-tests and found that there was only difference in alertness in males between before to after the test. Other variables did not differ between before and after the test.