心理学报 2019, Vol. 51, No. 5, 571–583
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2019.00571

第三方惩罚的神经机制：来自经颅直流电刺激的证据*

殷西乐[1,2] 李建标[3,4] 陈思宇[1] 刘晓丽[4] 郝 洁[5]

([1]浙江工商大学工商管理学院; [2]浙江工商大学浙商研究院; [3]浙江工商大学 MBA 学院, 杭州 310018)
([4]南开大学泽尔滕实验室, 南开大学滨海学院, 天津 300071) ([5]浙江工商大学财务与会计学院, 杭州 310018)

摘 要 第三方惩罚既是社会规范在群体得以维系的基石, 也是个体维护社会规范的体现。当前关注社会规范的神经研究大多基于第二方惩罚的独裁者或最后通牒实验框架, 缺乏对第三方维护社会规范过程中相关脑区活动的探索, 对这一过程的内在神经机制也不清楚。本文基于第三方惩罚的独裁者博弈框架, 对右侧背外侧前额叶区域(DLPFC)进行不同极性的经颅直流电刺激(tDCS), 同时依据第三方是否需要为其惩罚付出成本设计了零成本和有成本两个实验任务。结果发现, 第三方在零成本任务的情绪反应和惩罚显著受到 tDCS 设置的影响, 且阴极刺激显著提升了第三方的惩罚值, 这表明情绪机制对第三方惩罚有着重要影响。另外, 第三方在零成本和有成本任务中的惩罚差异在不同 tDCS 设置之间也存在显著差异, 这与第三方惩罚还受到自利机制影响的观点相符。本文率先为右侧 DLPFC 活动影响第三方惩罚提供了神经层面的证据, 且支持了第三方对社会规范的遵从与其负性情绪反应和自利加工密切相关的机制解释。

关键词 社会规范; 第三方惩罚; 背外侧前额叶; 经颅直流电刺激; 情绪
分类号 B845; B849: C91

1 前言

社会规范是反映社会成员共有价值观和信念的非正式制度(Elster, 1989), 是被群体成员广泛认可和遵从的行为准则和规范(Cialdini & Trost, 1998; 陈思静, 何铨, 马剑虹, 2015)。借助社会规范管控社会交往已经成为人类文明的主要标志之一(Ruff, Ugazio, & Fehr, 2013), 而社会规范在群体的维系离不开对规范违反行为进行的制裁和惩罚, 否则群体成员对社会规范的遵从会很快瓦解(Fischbacher, Gächter, & Fehr, 2001; Ruff et al., 2013)。这里的惩罚既可能来自第二方(其利益受到社会规范违反行为的直接影响), 也可以来自类似旁观者的第三方。例如, “路见不平, 拔刀相助”可看作是第三方维护社会规范的生动过程(陈思静等, 2015)。实验研究表明, 第二方和第三方都有惩罚社会规范违反行为的动机, 这些惩罚是社会规范得以施行的前提, 也是个体遵从社会规范的体现。然而, 如果只有第二方进行惩罚, 社会规范的影响将十分有限。因此, 第三方惩罚是社会规范得以大规模维系的基石和关键 (Fehr & Fischbacher, 2004; Fehr & Gächter, 2002)。

以 Fehr 为代表的学者最早把第三方惩罚与社会规范联系起来(陈思静等, 2015)。例如, Fehr 和 Fischbacher (2004)在独裁者博弈的基础上添加一个拥有惩罚权的第三方, 发现第三方存在强烈的社会规范遵从倾向：大约有 60%的第三方会付出成本去惩罚做出不公平分配的独裁者, 即便这些独裁者没

收稿日期: 2018-08-10
* 浙江省高校人文社会科学重点研究基地(浙江工商大学工商管理学科) (JYTgs20181107)、国家自然科学基金(71673152)资助。
通信作者: 李建标, E-mail: biaojl@126.com; 郝洁, E-mail: haojie@mail.zjgsu.edu.cn

有直接损害第三方的利益[1]。由于该框架下第三方的利益不受社会规范违反行为(不公平分配)的影响，其惩罚行为可以十分“干净”和“无污染”地体现社会规范所起的作用，因此第三方惩罚的独裁者博弈等成为考察个体社会规范遵从行为的经典框架。

第三方惩罚的行为实验虽然提供了个体遵从社会规范的直接证据，但难以从更深层解释惩罚动机的内在机制。换言之，我们对第三方维护社会规范的认知和情绪基础仍然知之甚少，自然也难以解释社会规范是如何形成的，以及其在特定场景下如何以及为什么会改变(Fehr & Fischbacher, 2004)。对个体行为最直接、最有说服力的证据来自人类的大脑。特别是，借助一些神经科学手段，可以找寻埋藏在行为规律背后的情绪或认知加工层面的神经证据。然而，学界对第三方惩罚即其维护社会规范的内在神经机制一直缺乏清晰的认识。

进一步，Jordan, Mcauliffe 和 Rand (2016)发现负性情绪是驱使第三方维护社会规范的重要动机：第三方自我汇报的愤怒情绪与其惩罚值显著相关。在人类的大脑区域中，右侧背外侧前额叶(dorsolateral prefrontal cortex, DLPFC)被认为是负责情绪和理性加工的重要脑区(Sellaro, Nitsche, & Colzato, 2016; 罗艺，封春亮，古若雷，吴婷婷，罗跃嘉，2013; 王益文等，2011)，因此，该区域活动应该与第三方惩罚密切相关。事实上，部分学者基于最后通牒博弈框架已经关注了右侧 DLPFC 在第二方的社会规范遵从中所起的作用。例如，Ruff 等人(2013)通过经颅直流电刺激(transcranial direct current stimulation, tDCS)发现提议者对社会规范的遵从与右侧前额叶(LPFC)的活动有关，Knoch 等人(2008)则发现，当阴极 tDCS 刺激抑制了右侧 DLPFC 后，回应者更容易接受不公平的提议。

然而，部分研究借助功能核磁共振(fMRI, functional magnetic resonance imaging)和 tDCS 技术发现，右侧 DLPFC 活动与第三方对不公平行为的惩罚没有显著关系(Civai et al., 2015; Corradi-Dell'Acqua et al., 2012)。例如，Corradi-Dell'Acqua 等人(2012)通过 fMRI 发现内侧前额叶(medial prefrontal cortex, MPFC)和右侧 DLPFC 只与第二方惩罚有关，而与第三方惩罚无关。Sellaro 等(2016)据此认为，DLPFC 活动只与那些受到不公平对待的第二方社会规范有关，而与普适性的社会规范遵从(第三方惩罚)无关。考虑到 DLPFC 活动对个体的情绪加工有着重要影响，如果 DLPFC 的活动不影响第三方的惩罚，这是否意味着第三方惩罚的生成和维系取决于情绪机制这一观点不成立呢？

本文认为，第三方的 DLPFC 活动不影响其社会规范遵从行为只是表象，原因在于 DLPFC 的活动不仅影响情绪机制，还通过自利机制(self-interest goals)对其惩罚行为产生了相反的影响。一方面，DLPFC 的活动会压制负性情绪(Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; 罗艺 等, 2013; 吴燕，周晓林，2012)，从而降低了第三方维护社会规范的动机；另一方面，DLPFC 同时抑制了个体的自利倾向(Knoch et al., 2008; Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006)，这相当于降低了第三方对惩罚成本的敏感性，从而提高了第三方维护社会规范的意愿。这两类机制相互冲突且影响幅度几乎同等重要，最终使得 DLPFC 活动没能从行为层面上改变第三方惩罚。换言之，本文认为第三方的社会规范遵从行为背后存在情绪和自利机制，且两类机制的作用方向相反。

依据这一观点，我们能够较好地解释 DLPFC 活动与第三方惩罚无关的已有研究困惑。在 Corradi-Dell'Acqua 等人(2012)和 Civai 等人(2015)的实验框架中，第三方本身是最后通牒博弈中的回应者，只不过他们的决策不影响自身和同组提议者的收益，而是代替下一组的回应者决策。这样，某回应者所惩罚的提议者并没有侵犯该回应者的利益，从而产生了第三方的概念。然而，这一设计使得第三方的惩罚不仅会降低违反规范的提议者的利益，也损害了与该提议者同组的回应者的利益，这一干扰项使得此类研究可能低估了社会规范的影响：由于第三方需要为其惩罚行为承担(心理)成本，此时自利和情绪机制的冲突掩盖了 DLPFC 在社会规范中起到的作用。

为验证本文研究观点，我们借鉴 Fehr 和 Fischbacher (2004)的第三方惩罚独裁者博弈框架，区分了第三方惩罚是否需要付出成本两类情形，这一设计避免了已有研究对两类机制的混淆。进一步，我们通过 tDCS 技术外生改变 DLPFC 脑区的活跃水平，继而观察第三方惩罚是否发生变化，通过情

[1] 该实验思路是，被试首先被分为独裁者、接受者或第三方三种类型之一，其中独裁者获得 100 点实验币并在自己和接受者之间任意分配，接受者只能被动接受独裁者的分配方案，不能影响任何人的收益。这一过程中第三方可以选择付出一定的成本去惩罚提出不公平提议的独裁者，规则是第三方自身每付出 1 点成本，独裁者的收益将被扣除 3 点。

绪和自利机制剖析了第三方对社会规范的遵从过程, 率先为第三方为何维护社会规范提供了神经层面的证据和机制解释。具体设计如下：遵循“组间(tDCS 刺激组) × 组内(惩罚的成本设置)”的实验设计思路, 首先依据实验被试所接受的刺激电极将被试随机分入阳极刺激组、阴极刺激组和虚拟刺激组; 其次, 三个刺激组的被试均需完成两个第三方惩罚的独裁者博弈实验任务, 其中任务 1 中被试作为第三方不需为其惩罚承担任何成本(零成本设置), 任务 2 中被试第则需要为其惩罚行为付出成本(有成本设置)。

在零成本设置中, 由于第三方无需为其对不公平提议的惩罚付出任何成本, 因此这一惩罚过程不涉及自利机制, 而主要受情绪机制的影响; 在有成本设置中, 第三方的惩罚同时受到情绪机制和自利机制的影响。这意味着, 通过分析第三方在有无成本设置中的惩罚值的差异, 可以考察自利机制的作用。因此, 上述实验设计使得本文能够区分情绪机制和自利机制在第三方惩罚过程中所起作用。

具体而言, 由于 DLPFC 的活动会压制负性情绪和自利加工过程, 本文预计：相比虚拟刺激组, 阴(阳)极刺激 DLPFC 释放(抑制)了第三方的负性情绪和自利考量, 从而提高(降低)了(1)被试在零成本设置中的惩罚值, (2)被试在两种成本设置中惩罚值的差异。如果(1)和(或) (2)得到实验结果支持, 就可以印证第三方惩罚受到情绪和(或)自利机制的影响。

综上, 当前关注社会规范的行为和神经实验研究为个体维护社会规范提供了有力证据。然而, 现有研究特别是神经实验研究大多基于第二方惩罚框架, 对第三方惩罚行为关注较少, 对其内在的作用机制也不清楚, 而后者对社会规范的维系更为重要。本文利用 tDCS 技术刺激右侧 DLPFC, 解析了情绪和自利机制在第三方惩罚中所扮演的作用, 回答了第三方为什么会遵从社会规范这一问题, 这不仅为行为科学的第三方惩罚及其社会规范遵从理论提供了直接的神经证据, 还有助于打开社会规范得以维系的机制黑箱。

2　实验设计

2.1　被试

实验共包括来自南开大学的 90 名本科和硕士研究生, 其中男性 41 人, 女性 49 人, 平均年龄 22 岁(年龄区间为 18~26 岁)。所有被试均为右利手且身体健康, 视力正常或矫正视力正常, 没有精神系统病史及脑部损伤病史, 也没有 tDCS 实验经历。实验符合 Helsinki 条款且获得所在实验室伦理委员会批准, 同时被试在参加实验时均已签署实验知情书。被试被随机分为三组：阳极刺激组、阴极刺激组和虚拟刺激组, 每组包括 30 人。有一个阴极刺激被试头皮阻抗过高, 未能进行稳定持续的电流刺激, 因此将其数据剔除后本文最终取得 89 个被试的有效数据。89 名有效被试中有 69 名被试专业为经管类(包括经济学、金融学、会计学和企业管理等), 占比为 77.53%。另外, 被试专业类型在不同刺激组中的分布比较接近：阴极、阳极和虚拟组分别包括 6、6 和 8 名非经管被试。

2.2　经颅直流电刺激技术(tDCS)

tDCS 是一种非侵袭性、利用微弱极化直流电(1~2 mA)调节大脑皮质神经细胞活动的技术, 它由阳极和阴极两个电极片构成(Filmer, Dux, & Mattingley, 2014; 甘甜 等, 2013; 甘甜, 石睿, 刘超, 罗跃嘉, 2018)。tDCS 刺激对大脑皮层兴奋性的影响具有极性特点, 阳极刺激(anodal stimulation)增强皮质兴奋性, 阴极刺激(cathodal stimulation)则相反(Civai et al., 2015; Nitsche & Paulus, 2001)。该技术的一个主要优点在于, 由于其可以外生改变受刺激大脑区域的活性, 因此如果个体行为在不同刺激组存在系统性差异, 我们就可以认定该大脑区域的活动(或与其有关的神经机制)对个体行为产生显著影响。因此, tDCS 可以建立受刺激的大脑区域和我们感兴趣的认知功能之间的因果关系, 具体可参阅 Filmer 等人(2014)和 Sellaro 等人(2016)的综述。

按照国际 EEG 10-20 系统的标准, tDCS 的目标电极放于右侧 DLPFC 所对应的 F4, 这是针对右侧 DLPFC 的 tDCS 研究常用的刺激位置(Li et al., 2017; Tremblay et al., 2014; Wang, Li, Yin, Li, & Wei, 2016; Ye, Chen, Huang, Wang, & Luo, 2015)。另外, 参考电极参照 Meiron 和 Lavidor (2013)、Harty 等人(2014)位于顶区的 Cz 位置[2]。刺激仪器采用德国 neuro Conn 公司的 DC-STIMULATOR, 刺激电极片

[2] 对于以右侧 DLPFC 为目标脑区的 tDCS 研究, 选取 Cz 作为参考电极位置是比较常见的选择(Ruff et al., 2013; Harty et al., 2014; Li et al., 2017)。头皮位置 Cz 对应于右侧和左侧中央沟的汇合, 相比其他位置(如眼眶), 该区域及其附近的脑区对认知活动的影响通常较小。特别是, Spitzer, Fischbacher, Herrnberger, Grön 和 Fehr (2007)的 fMRI 研究发现 Cz 附近的脑区没有在独裁者博弈中激活, Ruff 等(2013)也指出该区域作为 tDCS 参考电极不会干扰社会规范相关的神经活动。因此本文选取 Cz 作为参考电极位置。

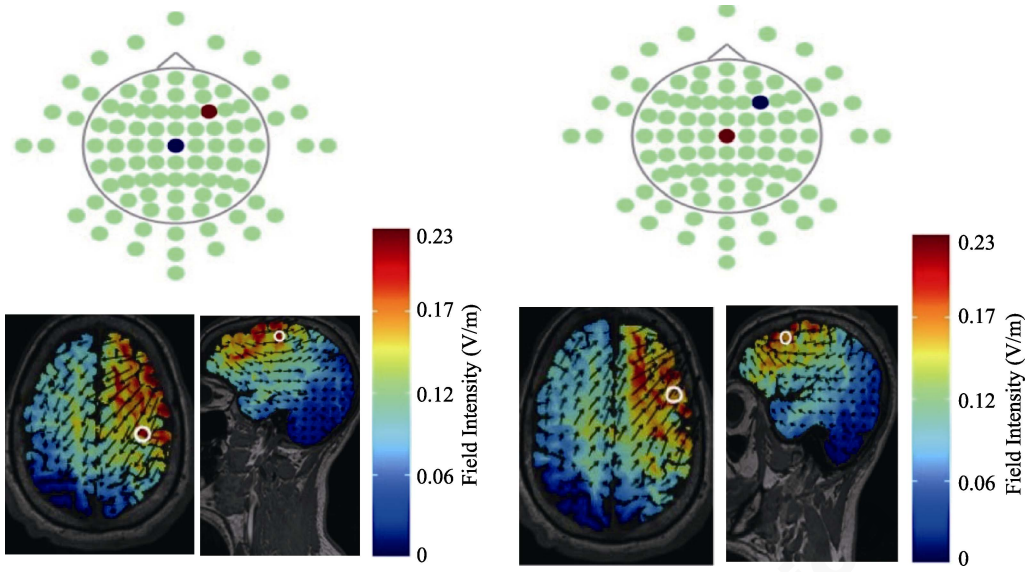

图 1 HD-Explore 软件绘制的电流模式展示(左图为阳极刺激，右图为阴极刺激)

大小为 5 × 7 = 35 cm^2，电流为 1 mA，有 15 秒上升和下降的缓冲时间，即刺激开始后电流强度逐步上升并在 15 秒后达到 1 mA。另外，本文使用 Soterix 医药公司(Soterix Medical Inc., New York)提供的 HD-Explore 软件，绘制了电极片位置并基于上述参数模拟了大脑区域接受刺激后的电流模式(图 1，箭头表示电流方向)。从中可以看出，无论是阳极刺激还是阴极刺激，均在右侧 DLPFC 区域产生了较高的场强(Field Intensity)，这表明刺激有效改变了该区域的活跃水平。

实验开始后，阴极组和阳极组的被试接受 15 分钟的电流刺激，而虚拟组被试所接受的刺激电流仅持续 30 秒。这一设计既保证了虚拟组被试的右侧 DLPFC 活动不因短暂电流刺激发生改变(Hummel et al., 2005; Willis, Murphy, Ridley, & Vercammen, 2015)，又使得被试在刺激开始阶段感到微痒等轻微刺激感，使其相信自己确实接受了电流刺激[3]。实验采用 offline 模式，即电流刺激结束后被试再参与实验任务，全部实验时间大约 40 分钟。实验报酬按照价值诱导原则(induced value)依据被试决策计算，平均收益为 48.6 元(包括 10 元被试费)。

2.3 实验任务与程序

实验采用 3(组间因素为 tDCS 刺激：阳极、阴极和虚拟) × 2(组内因素：任务 1、任务 2)的“组间+组内”设计。被试首先被随机分配到不同刺激组并接受不同极性的 tDCS 刺激，然后依次完成 2 个第三方惩罚的独裁者博弈实验任务，包括第三方无需为其惩罚付出成本的任务 1 (零成本设置)，以及第三方需要为其惩罚承担成本的任务 2 (有成本设置)。

任务 1 和任务 2 的设计参考 Fehr 和 Fischbacher (2004)，即通过第三方对独裁者的惩罚值考察其对社会规范的维护力度。实验中独裁者首先获得 100 实验代币(Game Dollar, G$)的初始禀赋，他需要决定将这 100 G$中的多少点分配给接受者，剩下的则留给自己。对于独裁者的分配方案，接受者只能选择接受。第三方在看到独裁者的分配方案后可以输入一个数值，规则是第三方每输入 1 G$，独裁者的收益将被扣除 3 G$。依据第三方输入的点数是否影响自身收益，实验分为零成本(任务 1)和有成本两个设置(任务 2)。在任务 1 中，第三方输入的数值不影响自身收益。在任务 2 中，第三方需要为自己输入的数值承担成本，即他每输入 1 G$，收益将被扣除 1 G$。为了控制学习效应的干扰(罗俊 等，2017)，有 50%的被试先完成任务 1 再完成任务 2，其他被试则完成任务 2 后再参加任务 1。

以零成本设置的任务 1 为例，其具体实验步骤如下。

1)所有被试均作为第三方参与实验，他们被告

[3] 本文 tDCS 实验采用双盲设计，包括被试和讲解实验说明的主持人在内均不知晓被试接受的是阳极刺激、阴极刺激还是虚拟刺激等信息。为实现这一目的，有两个主持人分别负责 tDCS 刺激和讲解实验说明。另外，该类实验之所以采用虚拟刺激设计而非让被试参加无脑电刺激的纯行为实验，是为了控制安慰剂效应(placebo effect)，该效应指被试认为自己接受的刺激或治疗对自己产生了效果或影响，但事实上该类刺激或治疗无效或并未执行(de la Fuente-Fernández et al., 2001; Li et al., 2017)。

知自己将在实验中充当观察者角色(C)并与位于另一个城市的东南大学行为决策实验室进行同步网络实验。

2)东南大学行为决策实验室包括 5 个小组, 每组都由独裁者(A)和接受者(B)两人组成。独裁者和接受者进行独裁者博弈实验, 其中独裁者对 100 G$进行分配, 而接受者只能接受独裁者的分配数额。

3)被试将依次看到 5 个组中的独裁者提出的分配方案。每观察一个小组的分配结果之前, 被试都会得到 50 G$。

4)每观察一个小组, 被试需要回答两个问题。①在看到独裁者分配结果后, 被试首先需要汇报对该分配结果的情绪反应, 该情绪反应是一个从非常愤怒到非常高兴的 5 分量表。然后, ②被试需要输入一个[0, 50]区间内的任意整数, 且被试每输入 1 G$, 独裁者的收益将被扣除 3 G$。被试无需为其输入的点数付出任何成本。

通常而言, 通过自我汇报(self-reported)的量表方式测度情绪需要小心对待, 因为被试在实验中自我汇报的情绪可能与其在现实决策情景下的情绪存在系统性偏差(Fehr & Fischbacher, 2004)。然而, 这一缺陷在本文 tDCS 研究中大大缓解, 原因在于, 不同刺激组中被试所经历的实验过程完全相同(且被试也不知晓自己接受了阴极、阳极或虚拟刺激中的哪一种), 因此被试自我汇报的情绪与现实情境存在的偏差(如果有的话)在不同刺激组中应该是一致的。由于本文主要关注不同刺激组中被试的情绪差异(例如相比虚拟刺激, 阴极刺激是否增强了负性情绪), 这意味着对不同刺激组情绪差异的比较会消除这一偏差的影响, 使得被试所表现出来的情绪差异只能由刺激这一因素所致。类似地, Ruff 等(2013)在其最后通牒博弈的 tDCS 研究中采用该方法测度了提议者对社会规范的信念, 以及其对回应者情绪反应的评估。

由于另一个实验区的独裁者和接受者的反应不是本文考察的重点, 他们事实上由计算机扮演, 但被试并不知晓这一点。另外, 实验中被试看到的独裁者的分配结果只有 0、30、50 三种可能, 参考王益文等(2014)和 Civai 等人(2015)的分类, 0 代表不公平的分配结果(unfair), 30 代表中等方案(mid-value), 50 则代表公平的分配方案(fair), 其中不公平的分配结果明显地违反了社会规范。

为确保被试相信上述内容的真实性, 实验采用了如下设计。第一, 被试在实验中被告知“你将与东南大学行为决策实验室进行同步网络实验, 该实验由汪敏达老师主持”这一比较场景化的信息, 而不是“你将与另一个实验区的参与人进行同步实验”等比较抽象的描述。依据认知集理论(cognitive set), 具象化的第一种表述方式更容易使被试联想到具体实验场景并在此基础上为接下来的任务做好准备(Ravizza & Carter, 2008; Rowe et al., 2007; Sakai, 2008), 因此能够降低被试对该表述方式真实性的怀疑程度。第二, 三种分配方案 0、30、50 中有两种结果随机重复呈现给被试, 这样被试一共观察 5 次分配结果。这一分配结果重复呈现的设计有助于减少实验结果的人为操纵痕迹[4]。第三, 实验中被试每观察一个小组, 计算机界面会显示“请等待第 *i* 组的 A 进行分配”, 且该界面持续时间为[5, 20]秒的随机数。这样, 被试在观察不同小组时, 会感知到不同组的独裁者的决策时间存在差异。第四, 被试在实验开始前阅读的实验说明中包含着独裁者和接受者的上机界面介绍(图 2 是独裁者界面), 这也增强了独裁者和接受者真实存在的可信性。

任务 2 过程与任务 1 类似, 不同之处在于任务 2 为有成本设置, 被试输入的惩罚点数将影响他的

[4] 如果本文没有重复设计, 被试将观察到三组独裁者的分配方案恰好不同(分别为 0、30 和 50), 这可能引起被试对实验真实性的怀疑, 因此我们将其中两种分配方案重复一次。与已有研究类似(Ruff et al., 2013; 王益文等, 2014), 本文采用伪随机方式控制随机设计对被试行为的影响。具体而言, 任务 1 和任务 2 中五组分配方案的出现次序进行伪随机排列, 规则是确保(0, 30, 50)三种分配方案至少出现一次且至多出现两次, 且两个任务的呈现顺序在被试间 ABBA 平衡。具体方式为, 第一场(session)实验中被试先进行零成本任务 1, 进行零成本任务 1 时计算机按照上述规则生成 5 组分配方案的呈现顺序 A, 随后在有成本任务 2 时按照上述规则随机生成 5 组分配方案的呈现顺序 B。在接来下的第二场实验中被试先进行有成本任务 2, 且其 5 组分配方案的呈现顺序为 A, 随后进行零成本任务 1, 且 5 组分配方案的呈现顺序为 B。也就是说, 计算机只在奇数场次实验(1, 3, 5,…, 29)的任务 1 和任务 2 生成 5 组分配方案的伪随机顺序, 而接来下的偶数场实验(2, 4, 6,…, 30)将任务呈现顺序颠倒, 但两个任务的 5 组分配方案的呈现顺序按照前一场实验的顺序(AB)进行。这一设计既控制了任务呈现的顺序效应, 又确保了零成本设置和有成本设置的 5 组分配方案的呈现顺序在总体样本上是一致的。进一步, 为了避免上述随机设计干扰 tDCS 效应, 我们每一场实验(session)招募三名被试, 且三名被试分别接受阴极、阳极和虚拟刺激。这样, 上述设计既减少了人为操作痕迹, 又确保了分配方案及其呈现顺序在不同 tDCS 组中完全相同。另外, 对于重复呈现的分配方案, 被试无需再次进行决策, 计算机将按照分配方案第一次呈现时他所做出的决策处理。

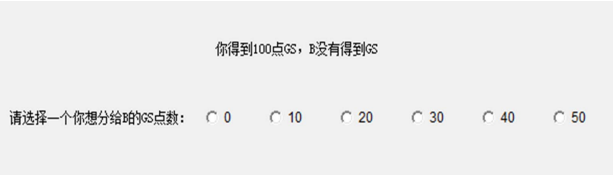

图 2 独裁者 A 的决策界面

收益，且任务 2 中被试无需汇报情绪反应[5]。被试在任务 1 和任务 2 中获得的 G$会在全部实验结束后按照 1 G$ = 0.08 元的比例兑换为现金。

实验任务采用组内设计，在任务 2 开始前，被试被告知他们与东南大学行为决策实验室的被试会被重新匹配分组。实验任务开始前被试还知晓位于东南大学行为决策实验室的其他被试不知晓本实验区被试的存在，这样设计的目的是避免独裁者策略性行为带来的干扰[6]。另外，实验采用 Z-tree 上机实验的方式(Fischbacher, 2007)，被试在完成任务 1 后才能看到任务 2 的具体描述和说明。实验说明采用抽象语境和中性语言，例如，被试在实验中被告知他们将充当 A、B 或 C 类型，而非独裁者等表述。同时，实验中避免采用“惩罚”等负面用语，而是使用“扣除”等中性表述替代。

在整个实验伊始，我们还测度了被试的社会规范信念(整个实验过程中只测量一次)。被试需要回答他们对社会规范的信念，即“你认为 A (独裁者)‘应该’向 B (接受者)分配多少 G$ (即 A 向 B 分配多少 G$是公平的)”。原因在于，除了情绪机制和自利机制，第三方惩罚还与他们对社会规范的信念和判断有关。这意味着，如果不同 tDCS 刺激组的被试对社会规范的理解存在系统性差异，那么本文研究结论的效力将会降低。虽然 Ruff 等(2013)发现通过 tDCS 刺激改变 DLPFC 的活动水平并不影响被试对社会规范的信念和判断，但鉴于这一因素的重要影响，本文希望进一步检验 tDCS 刺激是否显著改变第三方社会规范的信念，这有助于增强研究结论的说服力。

3 结果

3.1 第三方惩罚中的情绪反应

首先，从包括所有三个刺激组的全样本数据看，第三方情绪以及惩罚值与独裁者分配方案的公平性有关：分配方案越不公平，第三方的负性情绪越强，其在零成本设置中的惩罚值也越高(图 3)。具体而言，当独裁者给出不公平的分配方案时，第三方的负性情绪最强(均值为–0.89)，且该数值显著小于独裁者给出中等方案时的 0.15 (配对样本 t 检验，$t(88) = 9.93$, $p < 0.001$, $d = 1.23$)和公平的分配方案时的 1.35 ($t(88) = 15.64$, $p < 0.001$, $d = 2.77$)。由此带来的是，第三方在零成本设置时对不公平方案的惩罚值(21.64)显著高于中等方案时的 7.55 ($t(88) = 12.07$, $p < 0.001$, $d = 2.77$)和公平方案时的 1.58 ($t(88) = 13.15$, $p < 0.001$, $d = 1.83$)。上述结果不仅均在 1%水平显著，Cohen's d 效果量(effect size)也均大于 1.2 (一般认为 d 达到 0.5 效果为中，达到 0.8 即为效果大)，表明两组分布不重叠的部分至少在 62.2%以上。这一结果初步揭示了第三方存在维护社会规范的倾向，且这一行为与情绪机制存在密切的关系。

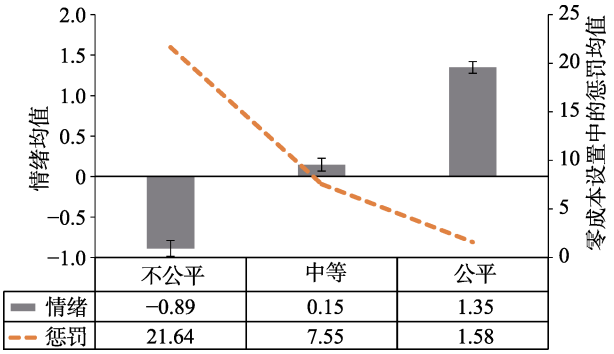

图 3 不同分配方案下第三方的情绪及其惩罚值(误差线为标准误)

其次，为检验第三方的负性情绪在不同刺激组是否存在显著区别，对第三方的情绪反应进行 3 (分配方案公平性：不公平、中等、公平) × 3 (tDCS：阴极、虚拟、阳极)的重复测量方差分析，结果显示 tDCS 设置的主效应显著($F(2, 86) = 4.29$, $p = 0.017$,

[5] 任务 2 无需汇报情绪反应的实验设计基于如下权衡：一方面，要求被试在任务 1 和任务 2 均汇报情绪反应可以确保两个任务的行为差别不受汇报与否的影响；另一方面，被试在有成本设置的任务 2 中汇报的情绪反应容易遭受“污染”，即其可能受到自利因素的干扰。就本文研究目的而言，我们希望相对“干净”地比较不同 tDCS 组被试的情绪反应是否存在差异进而考察第三方惩罚的情绪机制，因此更倾向于规避后者的影响。进一步，由于不同刺激组的被试面临的实验场景和流程完全相同，即便被试行为在任务 1 中受到情绪汇报的影响，这一偏差也会在对不同刺激组的结果比较中很大程度上得以消除。基于以上考虑，我们让被试只在零成本设置的任务 1 汇报情绪反应。

[6] 如果独裁者知晓有第三方存在，那么独裁者可能出于策略性的自利考量(即为了规避第三方的惩罚)而选择公平的选项，且独裁者的这一考量还受到第三方的惩罚是否存在成本的影响，这反过来可能会影响第三方在不同任务设置(任务 1 和任务 2)下的情绪和惩罚反应，从而对本文实验结果产生干扰。

偏 $\eta^2 = 0.091$), 交互效应也显著($F(4, 172) = 6.07$, $p < 0.001$, 偏 $\eta^2 = 0.124$)。这一结果表明对 DLPFC 进行 tDCS 刺激显著改变了第三方的情绪反应。

另外, 由图 3 可知当分配方案不公平时, 第三方产生了负性情绪(情绪均值为–0.89 < 0), 而当分配方案为中等或公平时, 第三方的情绪反应均值均大于零, 即产生了积极的正性情绪。为进一步检验分配方案公平性的影响, 接下来依据独裁者的三种分配方案分别进行方差分析, 描述性统计结果见表 1。

表 1　独裁者博弈中第三方情绪反应的描述性统计

方案公平性	tDCS	*n*	*M*	*SD*	ANOVA
不公平	阴极	29	–1.38	0.78	$F(2, 86) = 12.07$, $p < 0.001$, $\eta^2 = 0.219$
	虚拟	30	–0.97	0.72	
	阳极	30	–0.33	0.96	
	总体	89	–0.89	0.92	
中等	阴极	29	0.00	0.80	$F(2, 86) = 0.83$, $p = 0.438$
	虚拟	30	0.23	0.77	
	阳极	30	0.20	0.66	
	总体	89	0.15	0.75	
公平	阴极	29	1.52	0.57	$F(2, 86) = 1.43$, $p = 0.245$
	虚拟	30	1.30	0.70	
	阳极	30	1.23	0.73	
	总体	89	1.35	0.68	

当独裁者的分配方案不公平时, 单因素方差分析显示不同刺激组的情绪存在显著差异, 且 tDCS 主效应可以解释 21.9%的变异量, $F(2, 86) = 12.07$, $p < 0.001$, $\eta^2 = 0.219$[7]。事后比较(SNK)显示阴极组中第三方的情绪显著小于虚拟组(均值−1.38 < −0.97, $p < 0.05$), 阳极组中第三方的情绪均值为−0.33, 显著大于虚拟组($p < 0.05$)。另外, 单因素方差分析显示不同刺激组的情绪差异在中等水平的分配方案($F(2, 86) = 0.83$, $p = 0.438$)和公平的分配方案($F(2, 86) = 1.43$, $p = 0.245$)下均不显著。

上述结果表明, DLPFC 的激活抑制了不公平分配方案下第三方的负性情绪反应, 使得阳极组中第三方的负性情绪减弱, 而阴极组第三方的负性情绪增强。

需要指出, 本文实验同时考虑了独裁者提出的分配方案可能不公平、中等和公平等不同情形, 其中与本文研究主题最为密切地是不公平分配方案的情形。为了检验可能的无关因素(如实验控制)对实验结果的干扰, 本文还分析了分配方案中等和公平的情形。由于较为公平的分配方案不会明显违反社会规范, 因而第三方不会产生负性情绪并在其驱使下产生惩罚冲动。实验结果发现, tDCS 设置效应只在分配方案不公平时显著, 而在分配方案为中等和公平时不显著(后文 3.2 和 3.3 部分的结果与此相同), 这表明本文研究结果是稳健的。

3.2　零成本设置下第三方的惩罚值

在零成本设置下, 由于第三方无需为其对不公平提议的惩罚付出任何成本, 这一惩罚过程不会涉及自利加工, 而是主要受情绪机制的影响。因此, 本文预期如果 DLPFC 压制了负性情绪的生成(Knoch et al., 2006; Sanfey et al., 2003), 那么阴极刺激会增强被试的负性情绪体验, 使得被试在情绪冲动驱使下会付出更高的惩罚值(Fehr & Fischbacher, 2004)。阳极刺激则与之相反。

为检验第三方无需为其惩罚付出成本时其惩罚行为是否受到 tDCS 刺激的影响, 首先对零成本设置下第三方的惩罚进行 3 (分配方案公平性：不公平、中等、公平) × 3 (tDCS 设置：阴极、虚拟、阳极)的重复测量方差分析, 结果显示 tDCS 设置的主效应显著, $F(2, 86) = 4.69$, $p = 0.012$, 偏 $\eta^2 = 0.1$; 两因素的交互效应也显著, $F(4, 172) = 6.52$, $p < 0.001$, 偏 $\eta^2 = 0.132$。这一结果初步表明 tDCS 刺激显著改变了第三方的惩罚行为。下面依据独裁者的不同分配方案下分类讨论, 描述性统计结果见表 2。

表 2　独裁者博弈中第三方惩罚值的描述性统计(零成本设置)

方案公平性	tDCS	*n*	*M*	*SD*	ANOVA
不公平	阴极	29	29.59	9.88	$F(2, 86) = 9.29$, $p < 0.001$, $\eta^2 = 0.178$
	虚拟	30	19.37	13.74	
	阳极	30	16.23	13.19	
	总体	89	21.64	13.54	
中等	阴极	29	9.69	8.61	$F(2, 86) = 1.43$, $p = 0.246$
	虚拟	30	6.53	6.75	
	阳极	30	6.50	9.35	
	总体	89	7.55	8.35	
公平	阴极	29	1.76	9.28	$F(2, 86) = 0.07$, $p = 0.935$
	虚拟	30	1.83	9.14	
	阳极	30	1.17	3.13	
	总体	89	1.58	7.63	

首先, 对于不公平的分配方案, 单因素方差分析显示不同 tDCS 刺激组的第三方惩罚值存在显著

[7] 按照 Cohen (1988)给出的基准值参考, $\eta^2 = 0.01$、0.06 和 0.14 所对应的效果分别为小、中和大。

差异, $F(2, 86) = 9.29$, $p < 0.001$, $\eta^2 = 0.178$。进一步的事后两两比较检验(SNK)结果显示, 阴极组中第三方的惩罚额显著高于虚拟组(均值为 29.59 > 19.37, $p < 0.05$); 阳极组中第三方的惩罚额均值为 16.23, 与虚拟组差异不显著($p = 0.334$)。上述结果与第三方的情绪反应基本吻合。这表明, 当观察到社会规范被违反时, 第三方会产生负性情绪并在其影响下产生惩罚行为。一方面, 对第三方 DLPFC 进行 tDCS 阴极刺激释放了其负性情绪, 并使其在情绪的推动下产生了更高的惩罚水平。另一方面, 实验结果显示对第三方 DLPFC 进行 tDCS 阳极刺激压制了其负性情绪, 但没有显著降低其惩罚水平。

对于中等水平的分配方案, 单因素方差分析显示第三方的惩罚值在不同 tDCS 刺激组差异不显著, $F(2, 86) = 1.43$, $p = 0.246$; 公平的分配方案中也发现类似结果, $F(2, 86) = 0.07$, $p = 0.935$。

3.3 第三方在两种成本设置中的惩罚差异

当第三方需要为其惩罚承担成本时, 其惩罚行为不仅与情绪有关, 还受到自利目标的影响。因此, 第三方在零成本设置中的惩罚与其在有成本设置中惩罚的差值(以下用惩罚差异表述)可以测度自利机制对第三方惩罚的影响。由于 DLPFC 可以压制个体的自利倾向, 因此如果第三方惩罚过程中受到自利机制的影响, 本文预期当社会规范被违反时第三方的惩罚差异应该受到 tDCS 设置效应的影响, 且阳(阴)极刺激组中的惩罚差异显著低(高)于虚拟组。

为检验这一观点, 首先对第三方在两种成本设置下的惩罚差异进行 3 (分配方案公平性: 不公平、中等、公平) × 3 (tDCS 设置: 阴极、虚拟、阳极)的重复测量方差分析, 结果显示 tDCS 设置的主效应显著, $F(2, 86) = 4.04$, $p = 0.021$, 偏 $\eta^2 = 0.086$; 两个因素的交互效应也显著, $F(4, 172) = 3.95$, $p = 0.004$, 偏 $\eta^2 = 0.084$。

表 3 给出了惩罚差异在不同分配方案和 tDCS 刺激组中的描述性统计结果。首先, 当独裁者提出了不公平的分配方案时, 第三方在两种成本设置下的惩罚差异受到 tDCS 设置的显著影响, $F(2, 86) = 6.62$, $p = 0.002$, $\eta^2 = 0.133$。事后比较检验(SNK)结果显示, 阴极组中第三方在两种成本设置的惩罚差异显著高于虚拟组(均值为 21.90 > 12.67, $p < 0.05$); 阳极组中第三方在两种成本设置的惩罚差异均值为 10.10, 与虚拟组差异不显著($p = 0.451$)。

当独裁者提出了中等或公平的分配方案时, 第三方在两种成本设置下的惩罚差异均不受 tDCS 设置的显著影响($F(2, 86) = 1.70$, $p = 0.188$; $F(2, 86) = 0.08$, $p = 0.925$)。

表 3 独裁者博弈中第三方惩罚差异的描述性统计

方案公平性	tDCS	*n*	*M*	*SD*	ANOVA
不公平	阴极	29	21.90	13.10	$F(2, 86) = 6.62$, $p = 0.002$, $\eta^2 = 0.133$
	虚拟	30	12.67	12.92	
	阳极	30	10.10	13.17	
	总体	89	14.81	13.87	
中等	阴极	29	7.48	9.48	$F(2, 86) = 1.7$, $p = 0.188$
	虚拟	30	4.00	6.68	
	阳极	30	3.07	12.07	
	总体	89	4.82	9.74	
公平	阴极	29	1.76	9.28	$F(2, 86) = 0.08$, $p = 0.925$
	虚拟	30	1.33	9.28	
	阳极	30	0.97	2.65	
	总体	89	1.35	7.63	

本文还对第三方惩罚进行了 3 (公平水平: 不公平、中等、公平) × 3 (tDCS 设置: 阴极、虚拟、阳极) × 2 (成本设置: 零成本 VS 有成本)的重复测量方差分析, 结果显示 tDCS 设置($F(2, 172) = 3.96$, $p = 0.022$, 偏 $\eta^2 = 0.044$)和成本设置($F(1, 172) = 64.15$, $p < 0.001$, 偏 $\eta^2 = 0.272$)的主效应均显著, 且三个因素之间存在显著的交互效应($F(4.344) = 3.23$, $p = 0.013$, 偏 $\eta^2 = 0.036$), 这一结果与前述结果相一致。

最后, 单因素方差分析结果表明第三方对社会规范的信念在不同 tDCS 刺激组间的差异不显著, $F(2, 86) = 0.65$, $p = 0.524$。这一结果排除了 tDCS 刺激通过改变第三方对社会规范的信念和判断而改变其惩罚行为的可能性。我们还检验了两个实验任务的顺序安排是否存在顺序效应, 重复测量方差分析结果表明, 无论是第三方的情绪反应($F(1.87) = 0.63$, $p = 0.43$), 还是其在零成本设置下的惩罚值($F(1.87) = 0.30$, $p = 0.588$), 以及有成本设置下的惩罚值($F(1.87) = 0.54$, $p = 0.463$)均不受顺序设置的影响。除此之外, 我们还用独裁者分配额高于/低于第三方社会规范信念来定义“公平/不公平” (分配额小于社会规范信念定义为不公平, 否则为公平), 结果与本文已有结果一致: 当独裁者的分配方案不公平时, 零成本设置的单因素方差分析显示不同刺激组的情绪($F(2, 121) = 3.6$, $p = 0.03$, $\eta^2 = 0.096$)和惩罚值($F(2, 121) = 6.41$, $p = 0.002$, $\eta^2 = 0.056$)均存在显著差异, 且零成本和有成本设置的惩罚差异的 tDCS 效应也显著, $F(2, 121) = 3.65$, $p = 0.029$, $\eta^2 =$

0.057。当独裁者的分配方案公平时, 上述结果均不显著。因此, 本文研究结果是比较稳健的。

4　讨论

行为实验研究发现第三方观察到社会规范被违反时会产生愤怒等负性情绪, 并在情绪驱使下产生惩罚规范违反行为的内在动机(Fehr & Fischbacher, 2004; 陈思静, 马剑虹, 2011)。另外, 第三方的惩罚数量随着惩罚价格(成本)的上升而下降(范良聪, 刘璐, 梁捷, 2013), 因此维护社会规范的成本大小也会显著影响个体对社会规范的遵从水平。基于此, 本文预期第三方惩罚与情绪和自利机制密切相关。由于右侧 DLPFC 能够同时压制负性情绪(Sanfey et al., 2003; 罗艺 等, 2013; 吴燕, 周晓林, 2012)和自利倾向(Knoch et al., 2006; 2008), 本文借助 tDCS 技术改变右侧 DLPFC 的活跃水平, 进而考察情绪和自利机制如何影响第三方惩罚。实验结果发现, 第三方在零成本设置下的情绪和惩罚显著受到 tDCS 设置的影响, 且第三方在零成本和有成本设置中的惩罚差异在不同 tDCS 设置之间也存在显著差异。这一结果支持了第三方对社会规范的遵从同时受到自利和情绪机制影响的理论观点。

进一步, 本文认为自利和情绪机制的冲突掩盖了 DLPFC 在社会规范中起到的作用。为检验这一观点, 我们分析了任务 2 (有成本设置)实验结果, 发现与 Corradi-Dell'Acqua 等人(2012)和 Civai 等人(2015)的研究结果相一致：对第三方惩罚进行 3 (公平水平：不公平、中等、公平) × 3 (tDCS 设置：阴极、虚拟、阳极)的重复测量方差分析, 结果显示 tDCS 设置的主效应($F(2, 86) = 0.002$, $p = 0.998$)和两个因素的交互效应均不显著($F(4, 172) = 0.70$, $p = 0.592$)。进一步对 tDCS 设置的单因素方差分析结果也表明不同 tDCS 刺激组第三方的惩罚值在不公平的分配方案($F(2, 86) = 0.33$, $p = 0.723$)、中等分配方案($F(2, 86) = 0.48$, $p = 0.619$)和公平的分配方案下差异均不显著($F(2, 86) = 1.13$, $p = 0.327$)。这进一步表明当第三方维护社会规范需要付出成本时, 情绪机制与自利机制的相反作用导致了认知冲突, 而且二者对第三方惩罚行为的影响几乎同等重要, 使得两种机制权衡之下不同 tDCS 刺激组的差异不显著。

事实上, 现实生活中当付出较小的成本即能对规范违反行为施加惩罚时, 人类往往表现出较高的社会规范遵从水平, 但当施加惩罚显著影响切身利益时, 对社会规范的遵从就很可能被挤出。换言之, 当社会规范遵从需要付出成本, 情绪驱使下的维护社会规范的冲动与理性的自利思考会产生认知冲突, 且后者在一定程度上压制了情绪的影响。基于此, 本文认为第三方借助惩罚行为维护社会规范的内在机制是, 当看到某种规范违反行为时, 第三方的社会规范会被激活(Reno, Cialdini, & Kallgren, 1993)。这一过程不仅从认知层面唤醒其对社会规范概念和内容的认知, 还往往引起一系列负性情绪, 并驱使第三方产生惩罚社会规范违反行为的动机(陈思静, 马剑虹, 2011)。然而, 这一动机在付诸实施过程中还受到大脑关于自利理性的抑制(Zhou, Wang, Rao, Yang, & Li, 2014), 因此第三方的惩罚行为是情绪冲动和自利的理性思考权衡之下的结果。这一推测与基于第二方惩罚的神经科学实验的研究结果相符。例如, Feng, Luo 和 Krueger (2015)对基于最后通牒博弈框架的 fMRI 实验研究的元分析表明, 第二方惩罚的决策过程是大脑基于直觉的第一系统和深思熟虑的第二系统权衡之后的结果。大脑第一系统与情绪有关的活动首先产生惩罚规范违反者的冲动, 这一冲动与自利的目标产生冲突, 随后第二系统开始处理这一冲突, 并通过调节或压制情绪冲动和自利的冲突来达到最终的行为结果。

本文借助tDCS技术刺激右侧DLPFC探索了情绪和自利机制在第三方惩罚中的作用。这一分析的关键前提是, DLPFC 是否能够抑制个体的负性情绪反应和自利加工过程。情绪机制方面, Sanfey 等人(2003)和 Knoch 等人(2006)认为, DLPFC 通常与目标维护和执行控制等认知过程密切相关, 因此当不公平分配诱发第三方的负性情绪体验并产生惩罚冲动时, DLPFC 可以抑制这一情绪冲动。换言之, 虽然负性情绪的诱发和生成往往与负责情绪加工的脑岛(insula)、杏仁核(amygdala)等大脑区域而非 DLPFC 相关(Feinstein, Adolphs, Damasio, & Tranel, 2011; Gospic et al., 2011; 周平艳, 王凯, 李琦, 刘勋, 2012), DLPFC 却会在情绪冲动生成后对其加以抑制(罗艺 等, 2013; 吴燕, 周晓林, 2012)。特别是, Ochsner 等人(2004)和 Rêgo 等人(2015)分别借助 fMRI 和 tDCS 发现右侧 DLPFC 与负性情绪的控制(reduction、down regulation)有关。

另一方面, 个体自利加工和理性控制的过程与 DLPFC 密切相关。例如, Knoch 等人(2006; 2008)在最后通牒博弈中通过tDCS和TMS压制了回应者右侧 DLPFC 的活动后, 发现回应者对不公平提议

的拒绝率显著降低, 原因在于右侧 DLPFC 的活动压制了个体的自利倾向。进一步, Zhou, Wang, Rao, Yang 和 Li (2014)发现在分配比例相同的前提下(例如, 高分配金额设置下提议者得到 1600, 而回应者得到 400; 低分配金额设置下提议者得到 16, 而回应者得到 4), 响应者在高分配金额设置下的拒绝率明显更低(尽管个体对方案公平水平的判断相同), 且 fMRI 结果发现分配金额的高低调节了 DLPFC 的响应模式。

另外, 单因素方差分析的 SNK 事后比较检验结果表明, 对于不公平分配方案的 tDCS 刺激存在单边效应(polarity-specific effect): 零成本设置下的惩罚值、两种成本设置中的惩罚差异在阴极组时显著高于虚拟组, 但阳极组均不显著。事实上, 在与虚拟刺激比较时, 阴极和阳极二者之中只有一个显著的 tDCS 研究并不鲜见(Filmer et al., 2014; Jacobson, Koslowsky, & Lavidor, 2012)。例如, Filmer, Mattingley 和 Dux (2013)发现阴极刺激左侧后外侧前额叶(posterior lateral prefrontal cortex, pLPFC)能够显著提升个体在多任务决策中的反应速度, 阳极刺激则不显著。对本文单边效应的解释可以有两种。一是, 第三方的负性情绪和自利考量在其 DLPFC 活动处于正常水平时(虚拟刺激组)已经得到有效控制, 因此阳极刺激提升 DLPFC 的活跃水平未能进一步地改变两类机制的作用。二是, 情绪机制和自利机制对第三方惩罚的影响可能存在非线性的阈值效应: 当情绪和自利水平高于阈值(如虚拟刺激状态)时其对惩罚行为的影响比较明显, 而当其水平低于阈值时惩罚行为对它们的变化则相对不敏感。由于阴极刺激增强了负性情绪和自利考量, 而阳极刺激削弱了负性情绪和自利考量, 阈值效应可能导致了第三方惩罚在阴极刺激显著高于虚拟组而阳极刺激不显著的结果。

本文实验结果显示阳极刺激压制了第三方的负性情绪, 但没有显著降低其惩罚水平, 尤其是第三方在不公平分配方案下的情绪反应在阴极组为–1.38, 虚拟组为–0.97, 阳极组则为–0.33, 且阳极组显著高于虚拟组($p < 0.05$)。这一结果推翻了第一种解释, 表明第三方的负性情绪在虚拟刺激组中仍然比较明显, 且阳极刺激显著降低了其负性情绪反应。考虑到阳极刺激压制了负性情绪反应却没有显著降低其在零成本设置中的惩罚值, 因此第二种理论解释的可能性更高。

进一步研究可以从以下两点展开。第一, 本文重点关注了负性情绪在第三方惩罚中的作用, 但第三方对社会规范的遵从也可能受到正性情绪的影响。本文实验结果显示, 第三方在面临中等和公平的分配方案下情绪反应的均值大于零, 即总体上表现出了正性情绪。然而, 此类情景下情绪的 tDCS 设置效应不显著, 表明 DLPFC 的活动主要压制了负性情绪, 而对正性情绪的影响不明显。因此未来研究应关注影响情绪特别是正性情绪的其他脑区在第三方维护社会规范中的作用, 并且此类研究应该基于第三方帮助(而非惩罚)的实验框架。第二, 本文重点关注了与公平决策有关的社会规范, 未来研究还可以进一步检验情绪和自利机制是否适用于合作、保护环境等其他社会规范。这将有助于我们判定第三方对不同类型社会规范(如公平规范和合作规范)的遵从是存在着不同的工作原理, 还是有着共通的理论逻辑, 进而允许我们在实验场景变换之后, 预测第三方对社会规范的遵从是否以及如何发生变化。

5 结论

本文实验结果表明, DLPFC 是与第三方惩罚密切相关的重要脑区, 该区域的活动会显著改变第三方的负性情绪反应和自利加工过程, 进而影响第三方惩罚: 当看到违反社会规范的行为时, 第三方会受负性情绪驱使产生惩罚规范违反行为的冲动; 进一步, 如果对社会规范的遵从需要付出成本, 第三方自利的理性思考会削弱其情绪的冲动作用, 最终使得第三方惩罚即其对社会规范的遵从取决于其负性情绪和自利机制的权衡。

参 考 文 献

Chen, S. J., & Ma, J. H. (2011). Third-party punishment and social norm activation: The influence of social responsibility and emotion. *Journal of Psychological Science*, *34*(3), 670–675.

[陈思静, 马剑虹. (2011). 第三方惩罚与社会规范激活——社会责任感与情绪的作用. *心理科学*, *34*(3), 670–675.]

Chen, S. J., He, Q., & Ma, J. H. (2015). The influence of third-party punishment on cooperation: An explanation of social norm activation. *Acta Psychologica Sinica*, *47*(3), 389–405.

[陈思静, 何铨, 马剑虹. (2015). 第三方惩罚对合作行为的影响: 基于社会规范激活的解释. *心理学报*, *47*(3), 389–405.]

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 1–17). New York, NY, US: McGraw-Hill.

Civai, C., Miniussi, C., & Rumiati, R. I. (2015). Medial prefrontal cortex reacts to unfairness if this damages the self: A tDCS study. *Social Cognitive and Affective*

Neuroscience, 10(8), 1054–1060.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., & Fink, G. R. (2012). Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: An fMRI study. *Social Cognitive and Affective Neuroscience, 8*(4), 424–431.

de la Fuente-Fernández, R., Ruth, T. J., Sossi, V., Schulzer, M., Calne, D. B., & Stoessl, A. J. (2001). Expectation and dopamine release: Mechanism of the placebo effect in Parkinson's disease. *Science, 293*(5532), 1164–1166.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives, 3*(4), 99–117.

Fan, L. C., Lu, L., & Liang, J. (2013). The demand for the third party punishment: An Experimental Examination. *Economic Research Journal*, (5), 98–111.

[范良聪, 刘璐, 梁捷. (2013). 第三方的惩罚需求: 一个实验研究. *经济研究,* (5), 98–111.]

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63–87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137–140.

Feinstein, J. S., Adolphs, R., Damasio, A., & Tranel, D. (2011). The human amygdala and the induction and experience of fear. *Current Biology, 21*(1), 34–38.

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

Filmer, H. L., Dux, P. E., & Mattingley, J. B. (2014). Applications of transcranial direct current stimulation for understanding brain function. *Trends in Neurosciences, 37*(12), 742–753.

Filmer, H. L., Mattingley, J. B., & Dux, P. E. (2013). Improved multitasking following prefrontal tDCS. *Cortex, 49*(10), 2845–2852.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, 10*(2), 171–178.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters, 71*(3), 397–404.

Gan, T., Li, W. Q., Tang, H. H., Lu, X. P., Li, X. L, Liu, C., & Luo, Y. J. (2013). Exciting the right temporo-parietal junction with transcranial direct current stimulation influences moral intention processing. *Acta Psychologica Sinica*, *45*(9), 1004–1014.

[甘甜, 李万清, 唐红红, 陆夏平, 李小俚, 刘超, 罗跃嘉. (2013). 经颅直流电刺激右侧颞顶联合区对道德意图加工的影响. *心理学报, 45*(9), 1004–1014.]

Gan, T., Shi, R., Liu, C., & Luo, Y. J. (2018). Cathodal transcranial direct current stimulation on the right temporo-parietal junction modulates the helpful intention processing. *Acta Psychologica Sinica*, *50*(1), 36–46.

[甘甜, 石睿, 刘超, 罗跃嘉. (2018). 经颅直流电刺激右侧颞顶联合区对助人意图加工的影响. *心理学报*, *50*(1), 36–46.]

Gospic, K., Mohlin, E., Fransson, P., Petrovic, P., Johannesson, M., & Ingvar, M. (2011). Limbic justice--Amygdala involvement in immediate rejection in the Ultimatum Game. *PLoS Biology, 9*(5), e1001054.

Harty, S., Robertson, I. H., Miniussi, C., Sheehy, O. C., Devine, C. A., McCreery, S., & O'Connell, R. G. (2014). Transcranial direct current stimulation over right dorsolateral prefrontal cortex enhances error awareness in older age. *The Journal of Neuroscience, 34*(10), 3646–3652.

Hummel, F., Celnik, P., Giraux, P., Floel, A., Wu, W.-H., Gerloff, C., & Cohen, L. G. (2005). Effects of non-invasive cortical stimulation on skilled motor function in chronic stroke. *Brain, 128*(3), 490–499.

Jacobson, L., Koslowsky, M., & Lavidor, M. (2012). tDCS polarity effects in motor and cognitive domains: A meta-analytical review. *Experimental Brain Research, 216*(1), 1–10.

Jordan, J., Mcauliffe, K., & Rand, D. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics, 19*(4), 741–763.

Knoch, D., Nitsche, M. A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., & Fehr, E. (2008). Studying the neurobiology of social interaction with transcranial direct current stimulation—The example of punishing unfairness. *Cerebral Cortex, 18*(9), 1987–1990.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science, 314*(5800), 829–832.

Li, J., Yin, X., Li, D., Liu, X., Wang, G., & Qu, L. (2017). Controlling the anchoring effect through transcranial direct current stimulation (tDCS) to the right dorsolateral prefrontal cortex. *Frontiers in Psychology, 8*, 1079.

Luo, J., Ye, H., Zheng, H. L., Jia, Y. M., Chen, S., & Huang, D. Q. (2017). Modulating the activities of right and left temporo-parietal junction influences the capability of moral intention processing: A transcranial direct current stimulation study. *Acta Psychologica Sinica*, *49*(2), 228– 240.

[罗俊, 叶航, 郑昊力, 贾拥民, 陈姝, 黄达强. (2017). 左右侧颞顶联合区对道德意图信息加工能力的共同作用——基于经颅直流电刺激技术. *心理学报, 49*(2), 228–240.]

Luo, Y., Feng, C. L., Gu, R. L., Wu, T. T., & Luo, Y. J. (2013). The fairness norm in social decision-making: Behavioral and neuroscience studies. *Advances in Psychological Science*, *21*(2), 300–308.

[罗艺, 封春亮, 古若雷, 吴婷婷, 罗跃嘉. (2013). 社会决策中的公平准则及其神经机制. *心理科学进展, 21*(2), 300–308.]

Meiron, O., & Lavidor, M. (2013). Unilateral prefrontal direct current stimulation effects are modulated by working memory load and gender. *Brain Stimulation, 6*(3), 440–447.

Nitsche, M. A., & Paulus, W. (2001). Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. *Neurology, 57*(10), 1899–1901.

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D., & Gross, J. J. (2004). For better or for worse: Neural systems supporting the cognitive down-and up-regulation of negative emotion. *Neuroimage, 23*(2), 483–499.

Ravizza, S. M., & Carter, C. S. (2008). Shifting set about task switching: Behavioral and neural evidence for distinct forms of cognitive flexibility. *Neuropsychologia, 46*(12), 2924–2935.

Rêgo, G. G., Lapenta, O. M., Marques, L. M., Costa, T. L., Leite, J., Carvalho, S., ... Boggio, P. S. (2015). Hemispheric dorsolateral prefrontal cortex lateralization in the regulation of empathy for pain. *Neuroscience Letters, 594*, 12–16.

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology, 64*(1), 104–112.

Rowe, J. B., Sakai, K., Lund, T. E., Ramsøy, T., Christensen,

M. S., Baare, W. F. C., ... Passingham, R. E. (2007). Is the prefrontal cortex necessary for establishing cognitive sets? *Journal of Neuroscience, 27*(48), 13303–13310.

Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science, 342*(6157), 482–484.

Sakai, K. (2008). Task set and prefrontal cortex. *Annual Review of Neuroscience, 31*(1), 219–245.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300*(5626), 1755–1758.

Sellaro, R., Nitsche, M. A., & Colzato, L. S. (2016). The stimulated social brain: Effects of transcranial direct current stimulation on social cognition. *Annals of the New York Academy of Sciences, 1369*(1), 218–239.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., & Fehr, E. (2007). The neural signature of social norm compliance. *Neuron*, *56*(1), 185-196.

Tremblay, S., Lepage, J.-F., Latulipe-Loiselle, A., Fregni, F., Pascual-Leone, A., & Théoret, H. (2014). The uncertain outcome of prefrontal tDCS. *Brain Stimulation, 7*(6), 773–783.

Wang, G., Li, J., Yin, X., Li, S., & Wei, M. (2016). Modulating activity in the orbitofrontal cortex changes trustees' cooperation: A transcranial direct current stimulation study. *Behavioural Brain Research, 303*(4), 71–75.

Wang, Y., Wang Y., Lin C., Chen X., Yuan B., & Shen D. (2011). Modulation of conscientiousness on medial frontal negativity in negative emotions: An ERP study on ultimatum Game. *Scientia Sinica Vitae*, *41*(4), 320–331.

[王益文, 王钰, 林崇德, 陈雪莹, 袁博, 沈德立. (2011). 内侧额叶负波受负性情绪下责任感影响: 最后通牒任务的 ERP 研究. *中国科学: 生命科学, 41*(4), 320–331.]

Wang Y., Zhang Z., Zhang W., Huang L., Guo F., & Yuan S. (2014). Group membership modulates the recipient's fairness consideration in ultimatum game. *Acta Psychologica Sinica*, *46*(12), 1850–1859.

[王益文, 张振, 张蔚, 黄亮, 郭丰波, 原胜. (2014). 群体身份调节最后通牒博弈的公平关注. *心理学报, 46*(12), 1850–1859.]

Willis, M. L., Murphy, J. M., Ridley, N. J., & Vercammen, A. (2015). Anodal tDCS targeting the right orbitofrontal cortex enhances facial expression recognition. *Social Cognitive and Affective Neuroscience, 10*(12), 1677–1683.

Wu, Y., & Zhou, X. L. (2012). The context-dependency of fairness processing: Evidence from ERP study. *Acta Psychologica Sinica*, *44*(6), 797–806.

[吴燕, 周晓林. (2012). 公平加工的情境依赖性: 来自 ERP 的证据. *心理学报, 44*(6), 797–806.]

Ye, H., Chen, S., Huang, D., Wang, S., & Luo, J. (2015). Modulating activity in the prefrontal cortex changes decision-making for risky gains and losses: A transcranial direct current stimulation study. *Behavioural Brain Research, 286*, 17–21.

Zhou, P. Y., Wang, K., Li, Q., & Liu, X. (2012). Neural mechanisms of emotional modulation on memory. *Chinese Science Bulletin*, *57*(35), 3367–3375.

[周平艳, 王凯, 李琦, 刘勋. (2012). 情绪影响记忆的神经机制. *科学通报, 57*(35), 3367–3375.]

Zhou, Y., Wang, Y., Rao, L.-L., Yang, L.-Q., & Li, S. (2014). Money talks: Neural substrate of modulation of fairness by monetary incentives. *Frontiers in Behavioral Neuroscience, 8*, 150.

Neural mechanisms of third-party punishment: Evidence from transcranial direct current stimulation

YIN Xile[1,2]; LI Jianbiao[3,4]; CHEN Siyu[1]; LIU Xiaoli[4]; HAO Jie[5]

([1] *School of Business Administration, Zhejiang Gongshang University, Hangzhou* 310018, *China*)

([2] *Zheshang Research Institute, Zhejiang Gongshang University, Hangzhou* 310018, *China*)

([3] *MBA School, Zhejiang Gongshang University, Hangzhou* 310018, *China*)

([4] *Selten Laboratory, Binhai College, Nankai University, Tianjin* 300071, *China*)

([5] *School of Accounting, Zhejiang Gongshang University, Hangzhou* 310018, *China*)

Abstract

The social order of human societies is largely maintained by social norms. However, we still know little about the cognitive and emotional foundations that shape social norms, which makes it difficult, if not impossible, to understand how social norms are developed and maintained. Prior neural studies, which mainly perform second-party punishment based on the ultimatum framework, rarely explore the relevant brain areas as well as the neural mechanisms of third-party punishment driven by social norms. In the current study, we provide evidences that support the influences of two types of mechanisms (i.e., negative emotions and self-interest mechanisms) on social norms compliance of third parties at opposite directions. Meanwhile, right dorsolateral prefrontal area (DLPFC) is found to play a crucial role in this process.

In this study, we used transcranial direct current stimulation (tDCS) to investigate whether increasing or decreasing right DLPFC excitability influenced third-party punishment in a dictator game. Following an

experimental design of “between-subject (tDCS treatments: anodal, cathodal, sham) × within-subject (cost of punishment treatments: without cost, with cost)”, ninety participants were first randomly assigned to receive anodal, cathodal, or sham stimulation in 15 minutes. They then performed two dictator game tasks as third parties. In Task I (without cost) participants did not need to carry any costs for their punishment (none-cost task), while in Task II (with cost) they were required to pay for their punishment actions.

The results are given as follows. We first performed repeated measured ANOVA and one-way ANOVA to examine the effect of tDCS treatment (anodal, cathodal and sham) on emotion response. We found a significant main effect of tDCS on emotion response. Meanwhile, post-hoc analysis (SNK) showed that anodal stimulation decreased negative emotions while cathodal stimulation enhanced negative emotions. Second, the results of repeated measured ANOVA and one-way ANOVA showed a significant main effect of tDCS on punishment in the none-cost Task I, and post-hoc analysis (SNK) showed that cathodal stimulation significantly increased punishment while the effect of anodal stimulation was insignificant. Third, we also conducted repeated measured ANOVA and one-way ANOVA to test whether the difference of the punishment between the two tasks was affected by tDCS treatments. We found that the main effect of tDCS was significant. Moreover, post-hoc analysis (SNK) showed that the difference of punishment between the two tasks was significantly higher for cathodal stimulation than for sham stimulation, while the difference of punishment between the two tasks for anodal stimulation was insignificant compared to that of sham stimulation.

The present study provides one of the first neural evidences for the role of right DLPFC in third-parties’ social norms compliance. The results indicate that DLPFC, by affecting the processes of negative emotions and self-interest, is an important brain area of social norms compliance. When third parties face violations of social norms, their brains first release negative emotions that drive third parties to punish violators. Further, if third parties need to pay for their compliance with social norms, their rational goals about self-interest weaken negative emotional impulses. Finally, the compliance with social norms depends on the trade-offs between negative emotions and self-interest mechanisms.

Key words social norms; third-party punishment; dorsolateral prefrontal cortex (DLPFC); transcranial direct current stimulation (tDCS); emotion