

利他惩罚的认知机制和神经生物基础*

李 佳¹ 蔡 强² 黄禄华² 王念而² 张玉玲²

(¹ 东南大学学习科学研究中心儿童发展与学习科学教育部重点实验室, 南京 210096)

(² 北京大学心理系, 北京 100871)

摘 要 利他惩罚是指在团体中与他人合作, 并不惜花费个人代价去惩罚不合作者以维护群体规范的一种利他行为。这种行为所具有的利他特征, 可能是人类在漫长进化过程中形成的一种特定行为模式, 是人类本性的一个组成部分。本文结合已有研究, 总结了可用于研究利他惩罚的主要范式, 回顾了利他惩罚的进化背景, 从大量实证性文献中提炼出“不公平厌恶”及“心理理论和共情”两个维度来阐释引起利他惩罚行为的认知-情绪加工机制, 并整合了与其相关的脑神经证据, 最后对该研究领域未来的发展方向进行了展望。

关键词 利他惩罚; 不公平厌恶; 心理理论; 共情; 神经证据

分类号 B842

1 关于利他惩罚的概念

在人类漫长的演化历史中, 最初的合作秩序是基于自然选择的压力才迫使人类进化出有利于合作的偏好; 随着生产能力的提高, 自然施加于人类的选择压力开始减轻, 合作秩序不得不通过其他手段来维护, 强互惠者个人实施的利他惩罚就是其中之一(叶航, 汪丁丁, 罗卫东, 2005)。桑塔菲学派(SFI)认为, 强互惠(Strong Reciprocity)或者利他惩罚(altruistic punishment)是一种超越或突破“经济人”与“理性人”假说的人类行为模式, 是一种明显的亲社会行为(Gintis, Bowles, Boyd, & Fehr, 2003), 即超越“利己”动机, 在团体中, 宁可自己承担利益成本, 也会去惩罚那些不合作者以维护公正, 即使这些代价并不能得到预期的补偿(Fehr & Gächter, 2002)。利他惩罚行为本身不能直接增进群体中其他个体的生存适应性, 而是通过惩罚这种手段来维护合作秩序从而促进整个群体的平均适应性; 并通过群体中合作行为的增加而得到回报(Bowles & Gintis, 2004)。我们可以说,

有效的合作规范和秩序, 正是由人类作为生存竞争中最大的优势者, 通过有效地对背叛者进行惩罚实现的(Fehr & Gächter, 2002)。

由此可见, 界定利他惩罚需要具备三个主要特征: (1)这种惩罚不仅会给破坏群体规范的行为主体造成利益损失; (2)而且需要惩罚者付出一定代价; (3)这种行为从根本上在于推进社会福祉, 是一种利他行为。

2 研究利他惩罚的常用方法

在实验室情境中, 为多数学者用以研究利他惩罚的方法主要包括公共物品博弈(Public Goods Game, 简称 PGG) (Fehr & Gächter, 2000a; Fehr & Schmidt, 1999; Ledyard, 1995), 最后通牒博弈(Ultimatum Game, 简称 UG) (Camerer & Thaler, 1995; Güth & Tietz, 1990; Oosterbeek, Sloof, & van De Kuilen, 2004)以及第三方惩罚(Third-Party Punishment) (Bendor & Mookherjee, 1990; Charness, Cobo-Reyes, & Jiménez, 2008; Fehr & Fischbacher, 2004; Henrich et al., 2006)等。

2.1 公共物品博弈

PGG 是诱发利他惩罚的一个重要范式。博弈者向公共账户投资(投资额被增值), 进而获得最终的分红。就单个博弈者来说, 对公共账户的投资额越少, 其个人收益越趋近最大化; 同时, 这

收稿日期: 2011-04-28

*东南大学儿童发展与学习科学研究教育部重点实验室
开放研究基金资助课题(CDLS-2011-01)、国家自然科学基金项目(31100744)及浙江省自然科学基金项目(Y2110381)。

通讯作者: 李佳, E-mail: normilee@gmail.com

里还存在一个悖论,也就是说,一旦全体博弈者都去追求个体收益最大化,将直接导致公共账户失去投资额增值的机会而使整体收益趋近最小化,也就无从实现任何个体收益的最大化。在 PGG 中,对公共账户投资多的博弈者势必会受到投机者的拖累。博弈者在投资结束后有对其他博弈者进行惩罚(或有代价)的权利。大量的研究表明,人们会对那些投资比较少的人(搭便车者, free rider)进行惩罚,而且这种惩罚在大多数情况下可以促进博弈者的投资行为,即推动人际合作。

2.2 最后通牒博弈

该博弈对家分别称作分配方案的提议者(proposer)和反应者(responder)。由提议者提出金钱的分配方案,如果反应者接受此方案,那么博弈双方按照该方案进行分配;如果反应者拒绝这个提议,那么双方都得不到钱,从本质上导致双方利益同时受损,也可以看作是反应者对提议者的有代价惩罚。根据传统经济学“理性经济人”的假说,只要提议者分配给反应者的数额大于零,反应者就应该接受该方案;同时,提议者为了尽可能使自己利益最大化,则会提出使反应者收益无限趋近于 0 的提案。然而,实际研究发现,提议者分给反应者的平均配额在总额的 30% - 40% 之间;且少于总额 20% 的分配方案会被反应者拒绝(Kahneman, Knetsch, & Thaler, 1986; Thaler, 1988; Weg & Smith, 1993)。鉴于经典的单次 UG,反应者拒绝不公平分配方案可能是出于对社会规范和公平的考虑,而并非基于个体利益最大化的实现。

2.3 第三方惩罚

第三方惩罚作为一种可借助多种实验范式测量利他惩罚的实验情景,是指将惩罚的权利赋予不直接参与“交易”的第三方。作为旁观者,第三方不会在博弈互动中受到不公正的待遇,因此,第三方做出的(有代价)惩罚应该是出于纯粹对公平、公正的维护,而不牵涉个人利益在这种惩罚行为中的混淆。大量的研究证据表明,在不同的条件下(包括匿名条件,单次博弈),都存在相当一部分“强互惠者”,他们不仅会与其他个体互惠,而且不惜花费个人代价去惩罚那些破坏群体规范的个体。

3 利他惩罚的进化背景

人类所具有的合作特质使自然界中其他物种

只能望其项背(Fehr & Rockenbach, 2004; Wischniewski, Windmann, Juckel, & Brüne, 2009)。人类在大规模群体中会与非亲缘个体合作;人类合作也会发生在一次性交易的匿名个体之间,甚或发生在不能获得声誉收益的情况下。生物学家从自然选择出发,认为无论亲缘利他还是互惠利他从基因层面都体现了利己性——有利于个体适应性的生物性状才会在遗传中得以保存和进化,与个体适应性无益或有害的生物性状最终都会在遗传中丢失和湮没;认为利他行为不是一个进化稳定策略(Evolutionarily Stable Strategy)进而否定了生物的利他主义行为(Dawkins, 2006)。上个世纪 90 年代以来,随着实验经济学与行为博弈论的发展,经济学家发现,人类有相当一部分带有利他倾向的行为,却无法用亲缘理论和互惠理论解释(Dawkins, 2006; Rilling & Sanfey, 2011; Seymour, Singer, & Dolan, 2007)——能够给群体带来利益但却导致个体适应性降低的利他行为如何在严酷的生存竞争中对利己行为保持相对的遗传优势,并最终通过进化得以沿袭呢?

Bowles 和 Gintis (2004)认为,人类行为具有的这种利他特征,正是我们这个物种在漫长进化过程中形成的一种特定行为模式。当严酷的生存竞争迫使人类把合作规模扩展到血亲关系以外,而普遍存在的单次囚徒困境又无法为互惠行为提供条件时,由基因突变产生的强互惠或利他惩罚,可以侵入完全自私的人类群体;从而有效维护群内的合作规范,显著提高群内的生存竞争能力。接着,他们通过计算机仿真技术(Boyd, Gintis, Bowles, & Richerson, 2003),揭示了当不能维持利他性合作时,群间选择可以产生利他惩罚演化的可能:即在没有惩罚的条件下,群内适应会造成利他性合作的频次降低;而在加上惩罚条件后,能够在更大规模的群体中维持相当数量的合作行为。这体现了在自然选择和文化选择共同作用下的现代人类,既具有动物的自私性(经济理性),也具有社会性情感(即经济学所谓的非理性),正是这些社会性情感,例如同情、内疚等,在相当大的程度上决定了人类的利他行为。换句话说,我们不能把人类行为简单看作是追求物质效用最大化的工具。

在实验室操纵的 UG 中,当反应者拒绝提议者的分配方案时,就相当于对提议者进行了利他

惩罚(Fehr & Fischbacher, 2003; Seymour et al., 2007)。在有足够的惩罚威胁时,人们会变得更加慷慨,尤其是那些极为重视自我利益的人(Spitzer, Fischbacher, Herrnberger, Grön, & Fehr, 2007)。所以说惩罚可以对违反社会规范的行为起到威慑作用,促使人们做出公平的决策(Fehr & Fischbacher, 2004; Fehr & Gächter, 2000a, 2000b; Spitzer et al., 2007)。一方面,有效惩罚遏制了违规者的自利动机,从而避免了可能出现的负性后果;另一方面,有效惩罚会给惩罚者带来良好的感觉,并被认为是维护社会公正的正当策略,确保利他惩罚机制得到积极的正强化,从而肯定了族群发展中惩罚存在的意义,这与计算机模拟的人类社会利他惩罚演化的结果一致(Boyd et al., 2003)。当然,也有研究表明,过度的惩罚可能会破坏人的内在亲社会动机,从而不利于社会公平(Fehr & Rockenbach, 2003; Houser, Xiao, McCabe, & Smith, 2008; Li, Xiao, Houser, & Montague, 2009)。

4 利他惩罚的认知-情绪加工机制

实验室实验中,人们甚至在一次性(one-shot)交往中也会以一定的代价去惩罚不合作者(Fehr & Gächter, 2002; Ostrom, Gardner, & Walker, 1994);正如我们在日常生活中所感受到的:大多数人在发现那些违反社会规范的行为未得到惩罚时会感到不舒服,而一旦公正得以建立,他们就会感到轻松和满意(叶航,汪丁丁,罗卫东,2005)。

4.1 不公平厌恶

诚然自利是生物生存的本能(Dawkins, 2006),公平与互惠亦是社会进化的产物(Smith & Haakonssen, 2002)。自利与公平共同衍生出公平偏好和不公平厌恶。Fehr 和 Schimdt (1999)认为公平偏好表现为以自我为中心的不公平厌恶(Inequity Aversion)。不公平厌恶是指反对或抵制不公平结果的产生,比如当利益分配偏离了平等分配时,人们会体验到负效用,从而产生抗拒,会放弃一些利益(甚至自愿付出代价)以换取更为公平的结果。在经典的 PGG 中,博弈的最后一轮有 73%的个体(样本为 1042 人)拒绝投资,结果完全符合经济学中关于“理性人”的假设:通过搭便车来使个体的利益最大化。但实验后的调查却出乎人们意料,人们选择在 PGG 中拒绝或减少投资,

并非仅仅为了搭便车,大部分被试强调这样做是出于愤怒,希望通过减少投资的手段去惩罚那些纯粹的搭便车者,以维护公平的投资环境。为了验证被试们的解释,增加了被试有代价惩罚条件的后续实验结果表明,这种利他惩罚相当普遍,而且整体的投资水平也因此明显提高(Fehr & Gächter, 2002)。

Fehr 等人(2004; 1999; 2002)基于独裁者博弈(Dictator Game, 简称 DG)和囚徒困境博弈(Prisoners' Dilemma Game, 简称 PDG)的一系列研究发现,在 DG 中大部分的第三方会牺牲自己的物质利益来惩罚做出不公平提议的独裁者;在 PDG 中,46%的第三方会惩罚那些选择背叛合作方的个体。事实上,大量的行为博弈实验都证实了人们对公平这种社会理想的诉求和对社会规范的维护(Blount, 1995; Gintis, 2003; Güth, Schmittberger, & Schwarze, 1982)。

Polezzi 等人(2008)采用事件相关电位(Event-Related Potential, 简称 ERP)技术考察了反应者在 UG 中面对公平,亚公平和不公平分配方案时的脑电波形差异。结果发现,被试在亚公平和不公平条件下均产生了比公平条件下更负的反馈负波(Feed-back Negativity, 简称 FRN)。此成分与负性结果的加工有关(Oliveira, McDonald, & Goodman, 2007; Yasuda, Sato, Miyawaki, Kumano, & Kuboki, 2004)。这说明人们倾向于认为公平的提案是合理的,而不公平的提案则违反了反应者对公平的预期。此外,对公平感知的研究还发现了 MFN (Medial Frontal Negativity)与被试道德感正相关(Boksem & De Cremer, 2010)。

4.2 心理理论和共情

人际互动中,我们离不开对他人的知觉和理解。同样,在利他惩罚中,惩罚者的惩罚行为会受到惩罚结果有效性的影响。心理理论是指人们可以通过了解他人的信念、愿望和意图等心理状态来解释并据此推断他人行为的能力(Gallagher & Frith, 2003)。从 De Quervain 等人(2004)的实验中,我们能够看到,当信托人在信任游戏(Trust Game, 简称 TG)中背叛了投资人时,在有背叛意图无惩罚代价的条件下,全部被试都对背叛的信托人进行了惩罚;在有意图有惩罚代价的条件下,14 个被试中有 12 个对背叛者实施了惩罚;在有意图但只有象征性惩罚的条件下,实施惩罚的被试人数

锐减到不及一半数量;而在投资人的金钱损失不是信托人有意造成的条件下,投资人几乎没有惩罚的愿望,且惩罚强度相当低。这说明人们在做出惩罚行为时,不仅仅权衡了自身利益,还更多地考虑到博弈对家的意图。

也有学者指出,共情作为一种对他入情绪和精神状态理解以及对他人行为进行推测的能力(Singer, 2006),是自我与亲社会行为之间的一个重要中介变量(Davis, Kraus, & Ickes, 1997)。在自身利益不受博弈双方分配方案影响的第三方惩罚实验中,被试作为第三方通过观察博弈中分配方案提出者和接受者的行为,会对遭到不公平对待的受害者产生共情,去谴责或惩罚不公平的分配方案提出者(Buckholtz et al., 2008; Charness et al., 2008)。Falk, Fehr 和 Fischbacher (2005)指出对别人受到不公平待遇而感到愤怒的负性情绪和对规则破坏的不公平厌恶驱动了第三方惩罚,被试作为第三方愿意自付代价惩罚贪婪的规则破坏者来帮助不公平的受害者,以维持公正;研究结果也表明建立公正、对不公平行为进行惩罚能够诱发正性情绪(Civai, Corradi-Dell'Acqua, Gámez, & Rumiati, 2010; Singer et al., 2006)。共情是产生亲社会行为的一个普遍的动机基础(Hoffman, 2001)。来自神经方面的研究认为共情的神经基础可能在于杏仁核(amygdala, 负责社会性情绪加工,对于道德情景诱发的负性情绪敏感)及它与前额脑区底部皮层之间的联结(Buckholtz et al., 2008; Mitchell, 2010)。这体现了共情这种人的情绪体验与大脑边缘系统的密切相关,作为一种利他主义倾向,是人类本性的一个组成部分。

5 利他惩罚研究的主要神经证据

以牺牲个人利益来成全他人的“利他主义”并非仅在人类社会发生(Hamilton, 1963, 1964),但人类由于更经常地会在大规模群体中与非亲缘个体合作,使得人类社会存在不求外在回报,较其他高等动物的利他行为更为突出的真正利他行为;然而从演化的视角来看,这种行为却是使人费解的,因为合作的个体毕竟需要自己承担成本但却给非亲缘的成员带来利益——利他性惩罚可能是解释这种困惑的一个关键因素(Bowles & Gintis, 2004; Boyd et al., 2003; Fehr & Gächter, 2002)。那么利他惩罚背后的相关机制又是什么呢?神经经

济研究者(Fehr & Rockenbach, 2004)通过扫描在互动经济实验中被试做决策时的大脑,提供了利他惩罚的神经证据。研究结果支持个体情绪状态的神经表征决定人们的决策行为;同时,也强调了人们可以从互助合作、惩罚违规者这种行为中获得特殊的奖赏。

Sanfey, Rilling, Aronson, Nystrom 和 Cohen (2003)采用 UG 范式,在研究中发现,越不公平的分配方案越能引起被试双侧脑岛(bilateral insula)的激活,而该脑区通常是与恶心、疼痛、饥渴等引起的负性情绪相关。那些面对不公平分配方案,脑岛区激活程度越高的被试往往也更能拒绝不公平的分配方案。同时,研究者还发现人类对家给出的不公平分配方案较计算机对家给出的不公平分配方案会引起更强的脑岛激活水平,说明了社会性框架对于脑岛激活的重要性。不公平分配方案还引起了背外侧前额叶(dorsolateral prefrontal cortex, 简称 DLPFC)和前扣带回(anterior cingulate cortex, 简称 ACC)的激活。我们知道 DLPFC 是与目标维持和执行控制相关联的脑区且 ACC 主要负责认知冲突的感知。值得一提的是,在不公平分配方案的条件下,当脑岛的激活程度强于 DLPFC 的激活时,被试倾向于拒绝分配提议;当 DLPFC 的激活程度强于脑岛的激活时,被试往往会接受不公平分配方案。此外,腹内侧前额叶(ventromedial prefrontal cortex, 简称 vmPFC)和内侧眶额皮层(medial orbitofrontal cortex, 简称 mOFC)在惩罚决策中也发挥着至关重要的作用,与无代价惩罚条件相比,在有代价惩罚条件下,研究者发现了 vmPFC 更大程度的激活(De Quervain et al., 2004)。对该脑区的病理研究表明(Koenigs & Tranel, 2007),vmPFC 的损伤可能降低病患对奖赏的敏感性以及对负性情绪的控制,表现在 UG 中,会比对照组被试更多地拒绝分配者的不公平分配提议。

人们为什么会做出与“经济人”完全相反的行为,在没有物质补偿的情况下,不惜花费个人成本去惩罚那些违反合作规范的人?苏黎世大学国家经济实验室的学者猜测,如果不能从外界得到必要的激励,那么有代价的惩罚者只能从利他惩罚行为本身获得预期的满足。为了证实这个假设,他们使用 PET(Positron Emission Tomography)技术,考察了经济交易中对背叛者进行利他惩罚时

的脑神经机制(De Quervain et al., 2004)。实验采用TG范式,在惩罚阶段,设置了4个条件:(1)有意图有代价惩罚(intentional and costly,简称IC)(2)有意图无惩罚(intentional and free,简称IF)(3)有意图但象征性惩罚(intentional and symbolic,简称IS)(4)无意图有代价惩罚(no-intentional and costly,简称NC)。结果显示,被试在惩罚背叛者时激活了与奖赏相关的背侧纹状体(Dorsal Striatum),且背侧纹状体激活强的被试愿意花费更多的代价去对背叛者实施惩罚。以往在大鼠脑损毁性实验(Salinas & White, 1998)和灵长类动物的单一神经元记录实验(Apicella, Ljungberg, Scarnati, & Schultz, 1991; Hollerman, Tremblay, & Schultz, 1998)中,也均发现包括尾核(Caudate Nucleus)在内的纹状体(Striatum)的激活与奖赏信息和奖赏过程密切相关。此外,也有研究发现人类尾核的激活与奖赏过程相关(Delgado, Locke, Stenger, & Fiez, 2003; Delgado, Nystrom, Fissell, Noll, & Fiez, 2000; Delgado, Stenger, & Fiez, 2004; Knutson, Adams, Fong, & Hommer, 2001; Knutson, Westdorp, Kaiser, & Hommer, 2000; Koepp et al., 1998)。该研究支持了人们从惩罚违规者这种行为中获得满足的社会偏好模型(Camerer & Foundation, 2003; Fehr & Schmidt, 1999; Rabin, 1993),同时也力图阐释利他惩罚演化模型背后的相关机制。

6 问题和展望

利他惩罚作为一种捍卫公平、正义的亲社会行为在人类社会生活中并不鲜见,虽然我们不能从亲缘理论和互惠理论上对这种亲社会行为做出解释,但它的存在的确具有其适应生存的进化普适意义。公平,是人类追求理想社会的一个永恒主题。惩罚本身在现实中也执行着维系、捍卫公平的手段功能,利他惩罚更是从人性的道德高度揭示了人类对于公平的诉求和对于不公平的天生厌恶。但值得注意的是,这种似乎伴随人类进化成功留存下来的亲社会行为,仍有很多未解之谜等待我们去探究。比如,利他惩罚的心理根源尚存争议。到底利他惩罚是需要加入自我控制,应对社会规范的精细行为反应呢?还是仅仅作为由不公平厌恶的负性情绪带来的冲动性行为?人们究竟应该如何解读当身边人亦或自己产生“路

见不平,拔刀相助”行为时的真正动机?亲社会行为中的理智与情感如何进行较量?自利和利他是否真的完全相悖?同时,就在实验室开展社会决策研究所使用的方法本身而言,如何营造、模拟出趋近真实的情景,尽可能排除被试在实验过程中可能产生的社会赞许迎合和霍桑效应;研究结果在现实生活中能以多大程度予以推广以及如何解读被试间的个体差异……这些都是研究者不能忽视的问题。中国自古就有“不患贫而患不均”的说法,对利他惩罚的研究不仅在宏观框架下对公平、合作问题的探讨有重大的现实意义,而且新技术和多学科的交叉介入(Crockett, Clark, Lieberman, Tabibnia, & Robbins, 2010; Tabibnia, Satpute, & Lieberman, 2008)也必定有益于我们通过神经机制方面的证据来深化对这些问题的理解。

随着神经科学技术和社会认知神经科学的兴起,客观上推进了对该问题在神经层面的研究。但从可供采用的技术手段来说,研究诸如利他惩罚这种带有复杂的社会和文化内容的行为,并不能仅止于运用目前流行的研究手段(脑成像)得出一些脑区的激活;正如核磁技术本身,虽然具有良好的空间分辨率,但是时间分辨率较差(Logothetis, 2008),并不利于我们在时间进程上的认知分析。针对这个技术本身的缺陷,目前有两种方法可以在某种程度上给予弥补,其一是把每个认知过程分步呈现。比如我们在做赌博任务的时候,不同时呈现两张牌的结果,而是先呈现一张牌,在空余数秒后再呈现另一张牌。这种系列呈现的方式在一定程度上可以简化繁杂的认知过程,让我们对大脑上的激活有更清晰的解释,但必须指出的是,这是以牺牲生态效度为代价的。另外一种方法,现在尚处雏形——fMRI与ERP的结合(Horovitz, Skudlarski, & Gore, 2002; Iidaka, Matsumoto, Haneda, Okada, & Sadato, 2006),这样可以在保证空间分辨率的情况下,结合ERP手段确定认知的时程顺序(即在每个时间窗进行ERP源定位分析)。

其次,fMRI的原理是基于利用人体脱氧血红蛋白和含氧血红蛋白的不同磁性,间接测量某个脑区的代谢,方法上具有一定的间接性,这也是每个利用fMRI做神经成像的工作者需要注意的。此外,社会认知神经研究中,采用的血氧动力函

数在很大程度上借鉴了视觉皮层的血氧动力函数 (Huettel, Song, & McCarthy, 2004), 是否这个血氧动力函数可以普适在全脑也亟待我们更多的研究结果去验证。

最后, 社会认知是一个高级过程, 人类认知活动的复杂性决定了其映射到神经层面必定也是多个脑区联动的结果, 条件间可能会同时激活几个脑区, 所以在解释上会存在一些推测, 同时会存在一些实验更多地像是事后解释而非事前预测 (Mateo, Cabanis, Loebell, & Krach, 2011)。因此, 就目前对社会认知的研究现状而言, 单纯将某个认知活动与某个特定脑区僵化联结起来都不是审慎的做法 (Chiao et al., 2010)。这也是研究者需要在方法问题认识上保持警醒的地方。但是正是因为社会认知神经刚刚起步, 研究还比较粗糙, 所以才需要我们进行更加细化地研究, 设计更加巧妙的社会实验, 反复检验推敲各种可能与理论假设。同时, 一些脑损伤病人和对照组研究结论的吻合 (Greene, 2007) 也为我们提供着强有力的证据, 足以使我们相信, 社会认知神经科学尽管是一门新兴学科, 就像物理学刚起步一样, 虽然有诸多问题, 总体还在向一门趋向成熟的学科迈进。

参考文献

- 叶航, 汪丁丁, 罗卫东. (2005). 作为内生偏好的利他行为及其经济学意义. *经济研究*, 26(8), 84–90.
- Apicella, P., Ljungberg, T., Scarnati, E., & Schultz, W. (1991). Responses to reward in monkey dorsal and ventral striatum. *Experimental Brain Research*, 85(3), 491–500.
- Bendor, J., & Mookherjee, D. (1990). Norms, third-party sanctions, and cooperation. *Journal of Law, Economics, & Organization*, 6, 33–63.
- Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63(2), 131–144.
- Boksem, M. A. S., & De Cremer, D. (2010). Fairness concerns predict medial frontal negativity amplitude in ultimatum bargaining. *Social Neuroscience*, 5(1), 118–128.
- Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology*, 65(1), 17–28.
- Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6), 3531–3535.
- Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., et al. (2008). The neural correlates of third-party punishment. *Neuron*, 60(5), 930–940.
- Camerer, C. F., & Foundation, R. S. (2003). *Behavioral game theory: Experiments in strategic interaction* (Vol. 544). Princeton, NJ: Princeton University Press.
- Camerer, C. F., & Thaler, R. H. (1995). Anomalies: Ultimatums, dictators and manners. *The Journal of Economic Perspectives*, 9(2), 209–219.
- Charness, G., Cobo-Reyes, R., & Jiménez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior & Organization*, 68(1), 18–28.
- Chiao, J. Y., Hariri, A. R., Harada, T., Mano, Y., Sadato, N., Parrish, T. B., et al. (2010). Theory and methods in cultural neuroscience. *Social Cognitive and Affective Neuroscience*, 5(2-3), 356–361.
- Civai, C., Corradi-Dell'Acqua, C., Gamer, M., & Rumiati, R. I. (2010). Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. *Cognition*, 114(1), 89–95.
- Crockett, M. J., Clark, L., Lieberman, M. D., Tabibnia, G., & Robbins, T. W. (2010). Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. *Emotion*, 10(6), 855–862.
- Davis, M. H., Kraus, L. A., & Ickes, W. (1997). Personality and empathic accuracy. *Empathic Accuracy*, 5, 144–168.
- Dawkins, R. (2006). *The selfish gene*. Oxford: Oxford University Press.
- De Quervain, D. J. F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science*, 305(5688), 1254–1258.
- Delgado, M. R., Locke, H. M., Stenger, V. A., & Fiez, J. A. (2003). Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cognitive, Affective, & Behavioral Neuroscience*, 3(1), 27–38.
- Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., & Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, 84(6), 3072–3077.
- Delgado, M. R., Stenger, V. A., & Fiez, J. A. (2004). Motivation-dependent responses in the human caudate nucleus. *Cerebral Cortex*, 14(9), 1022–1030.
- Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, 73(6), 2017–2030.
- Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of

- social norms. *Human Nature*, 13(1), 1–25.
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785–791.
- Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63–87.
- Fehr, E., & Gächter, S. (2000a). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4), 980–994.
- Fehr, E., & Gächter, S. (2000b). Fairness and retaliation: The economics of reciprocity. *The Journal of Economic Perspectives*, 14(3), 159–181.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137–140.
- Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928), 137–140.
- Fehr, E., & Rockenbach, B. (2004). Human altruism: Economic, neural, and evolutionary perspectives. *Current Opinion in Neurobiology*, 14(6), 784–790.
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817–868.
- Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences*, 7(2), 77–83.
- Gintis, H. (2003). Solving the puzzle of prosociality. *Rationality and Society*, 15(2), 155–187.
- Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, 24(3), 153–172.
- Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Update*, 11(8), 322–323.
- Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367–388.
- Güth, W., & Tietz, R. (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology*, 11(3), 417–449.
- Hamilton, W. D. (1963). The evolution of altruistic behavior. *American Naturalist*, 97, 354–356.
- Hamilton, W. D. (1964). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1), 1–16.
- Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science*, 312(5781), 1767–1770.
- Hoffman, M. L. (2001). *Empathy and moral development: Implications for caring and justice*. Cambridge: Cambridge University Press.
- Hollerman, J. R., Tremblay, L., & Schultz, W. (1998). Influence of reward expectation on behavior-related neuronal activity in primate striatum. *Journal of Neurophysiology*, 80(2), 947–963.
- Horovitz, S. G., Skudlarski, P., & Gore, J. C. (2002). Correlations and dissociations between BOLD signal and P300 amplitude in an auditory oddball task: A parametric approach to combining fMRI and ERP. *Magnetic Resonance Imaging*, 20(4), 319–325.
- Houser, D., Xiao, E. T., McCabe, K., & Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior*, 62(2), 509–532.
- Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional magnetic resonance imaging* (Vol. 1). Sunderland, MA, Sinauer Associates.
- Iidaka, T., Matsumoto, A., Haneda, K., Okada, T., & Sadato, N. (2006). Hemodynamic and electrophysiological relationship involved in human face processing: Evidence from a combined fMRI-ERP study. *Brain and Cognition*, 60(2), 176–186.
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *The Journal of Business*, 59(4), 285–300.
- Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, 21(16), RC159.
- Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). fMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage*, 12(1), 20–27.
- Koenigs, M., & Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *The Journal of Neuroscience*, 27(4), 951–956.
- Koepp, M. J., Gunn, R. N., Lawrence, A. D., Cunningham, V. J., Dagher, A., Jones, T., et al. (1998). Evidence for striatal dopamine release during a video game. *Nature*, 393(6682), 266–267.
- Ledyard, J. O. (1995). *Public goods*. Princeton: Princeton University Press.
- Li, J., Xiao, E. T., Houser, D., & Montague, P. R. (2009). Neural responses to sanction threats in two-party economic exchange. *Proceedings of the National Academy of Sciences*, 106(39), 16835–16840.
- Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, 453(7197), 869–878.
- Mateo, M. M., Cabanis, M., de Echeverría Loebell, N. C., & Krach, S. (2011). Concerns about cultural neurosciences: A critical analysis. *Neuroscience & Biobehavioral Reviews*, 36, 152–161.
- Mitchell, D. G. V. (2010). The nexus between decision

- making and emotion regulation: A review of convergent neurocognitive substrates. *Behavioural Brain Research*, 217, 215–231.
- Oliveira, F. T. P., McDonald, J. J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: Expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience*, 19(12), 1994–2004.
- Oosterbeek, H., Sloof, R., & van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2), 171–188.
- Ostrom, E., Gardner, R., & Walker, J. (1994). *Rules, games, and common-pool resources*. Ann Arbor: University of Michigan Press.
- Polezzi, D., Daum, I., Rubaltelli, E., Lotto, L., Civai, C., Sartori, G., et al. (2008). Mentalizing in economic decision-making. *Behavioural Brain Research*, 190(2), 218–223.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5), 1281–1302.
- Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, 62, 23–48.
- Salinas, J. A., & White, N. M. (1998). Contributions of the hippocampus, amygdala, and dorsal striatum to the response elicited by reward reduction. *Behavioral Neuroscience*, 112(4), 812–826.
- 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.
- Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nature Reviews Neuroscience*, 8(4), 300–311.
- Singer, T. (2006). The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. *Neuroscience & Biobehavioral Reviews*, 30(6), 855–863.
- Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, 439(7075), 466–469.
- Smith, A., & Haakonssen, K. (2002). *The theory of moral sentiments*. Cambridge: Cambridge University Press.
- Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., & Fehr, E. (2007). The neural signature of social norm compliance. *Neuron*, 56(1), 185–196.
- Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness: Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, 19(4), 339–347.
- Thaler, R. H. (1988). Anomalies: The ultimatum game. *The Journal of Economic Perspectives*, 2(4), 195–206.
- Weg, E., & Smith, V. (1993). On the failure to induce meager offers in ultimatum game. *Journal of Economic Psychology*, 14(1), 17–32.
- Wischniewski, J., Windmann, S., Juckel, G., & Brüne, M. (2009). Rules of social exchange: Game theory, individual differences and psychopathology. *Neuroscience & Biobehavioral Reviews*, 33(3), 305–313.
- Yasuda, A., Sato, A., Miyawaki, K., Kumano, H., & Kuboki, T. (2004). Error-related negativity reflects detection of negative reward prediction error. *Neuroreport*, 15(16), 2561–2565.

Neurobiological Basis and Cognitive Mechanism of Altruistic Punishment

LI Jia¹; CAI Qiang²; HUANG Lu-Hua²; WANG Nian-Er²; ZHANG Yu-Ling²

(¹ Key Laboratory of Child Development and Learning Science (Ministry of Education), Southeast University, Nanjing 210096, China)

(² Department of Psychology, Peking University, Beijing 100871, China)

Abstract: People often punish other individuals who behave negatively or uncooperatively to promote cooperation even when it comes at personal cost, and when there is no expectation of receiving any material returns. We define this behavior as altruistic punishment. Here, we review the summation of experimental paradigm, process of altruistic punishment evolution and generalizes inequity aversion, theory of mind and empathy to be key factors which induce the behaviors of altruism punishment from current empirical literatures. In addition, altruism punishment-related brain basis has been reviewed and the review also highlights this field's possible future.

Key words: altruism punishment; inequity aversion; theory of mind; empathy; neural evidence