

公共物品困境中惩罚的形式与作用*

陈欣¹ 赵国祥¹ 叶浩生²

(¹ 河南大学教育科学学院, 心理与行为研究所, 开封 475004)

(² 广州大学心理与脑科学研究中心, 广州 510006)

摘要 公共物品困境中对惩罚的操纵表明: 惩罚搭便车可以维持群体内的合作行为。早期研究表明, 惩罚的设置有助于建立社会合作规范, 有代价惩罚、利他惩罚和第三方惩罚等惩罚形式为间接互惠提供了合理的解释。随着实证研究的进一步深入, 惩罚表现出新的形式, 如自私惩罚和反社会惩罚, 被惩罚者的威胁报复将使合作者不再继续惩罚搭便车行为, 导致合作行为的减少直至消失。正当性理论、道德情感理论和文化基因协同进化理论表明惩罚的存在具有一定的必要性和价值, 惩罚的作用存在框架效应。本文介绍了一些有关的中介变量, 如利他、信任、声誉、文化、期望等, 可以对惩罚作用的积极面和破坏性予以适当的解释, 也为合作的心理机制研究提供新的视角。

关键词 公共物品困境; 搭便车; 惩罚; 合作

分类号 B849:C91

1 引言

公共物品困境(public-goods dilemmas)是一种社会困境(social dilemmas), 日常生活中表现为“众人拾柴火焰高”, 是一种“我为人人, 人人为我”的公共物品供给与消费方式。在公共物品博弈中(public-goods games), N 个参与者拥有一定数额的实验代币(experimental currency units), 给所属的公共账户进行捐献(voluntary contributions mechanism), 公共账户的代币增值后平均分配给该账户的每个成员。参与者捐献得越多, 公共账户增值得越多, 个人收益也越多。若是参与人都不捐献, 那么公共账户没有收入来源, 无法增值产生收益, 大家都一无所获。少数参与人不捐献或者是少捐献, 也能从这种平均分配的机制中获利, 并且个人拥有的代币将多于那些一直高捐献的人。公共物品博弈将捐献数额作为“合作”(cooperation)的指标, 不捐献任何东西, 并且享受

由他人支撑的公共账户的行为被称之为“搭便车”(free-riders)。

在公共物品博弈中, 无论其他参与者做出什么决策, 搭便车使个人收益最大, 对个人而言是一个最优策略, 只是个人最优策略的集合将产生一个最坏的结果, 收益为零。公共物品困境呈现出一幅个人理性和群体理性存在冲突的场景, 追求个人利益最大化导致对所有人都不利的局面(Fehr & Schmidt, 1999; Greenberger, Miceli, & Cohen, 1987)。传统经济学认为公共物品困境避免不了搭便车, 预测“一个理性人捐献的代币数将为零, 公共系统无以为继”(Liebrand, Messick, Yamagishi, & Hayashi, 1996; Yamagishi & Sato, 1986)。行为经济学发现, 公共物品中确实存在搭便车, 但没有传统经济学预测的那么悲观。在重复遇到陌生人的实验方案中, 在实验初期参与人的捐献为个人拥有代币的一半, 随着实验的继续, 捐献水平逐渐下降, 慢慢地接近传统经济学的预测, 逐渐降低为零(Rubin, 1986)。间接互惠(indirect reciprocal)的观点能部分解释捐献衰退的现象, 公共物品捐献是一个自愿的“我助人人, 人人助我”的互惠过程, 搭便车破坏了这种潜在的互惠规范, 高捐献者对搭便车不满意, 逐渐丧失

收稿日期: 2013-05-13

* 河南省教育厅人文社会科学(2012-QN-041)和河南省软科学(132400410294)资助。

通讯作者: 叶浩生, E-mail: yehaosheng0817@163.com

捐献的意愿, 捐献呈盘旋下降趋势, 直至为零 (Skatova & Ferguson, 2013)。

公共物品供应的崩溃始于搭便车的逐步蔓延, 如何保证公共物品捐献的继续呢? 经济学、社会学、生物学和心理学从各自学科视角展开研究, 发现影响个体捐献的诸多因素中, 惩罚 (sanctions or punishment) 是有效的举措之一。大量研究表明, 惩罚有效地减少了公共物品中的搭便车, 增加捐献水平, 维持群体内的合作 (de Quervain et al., 2004; Skatova & Ferguson, 2013)。捐献数额作为因变量, 反映“合作”的程度, 惩罚的作用引起了研究的广泛兴趣, 本文关注两类问题: 一是惩罚的不同形式对合作水平的影响; 二是惩罚为什么能维持合作? 惩罚作为自变量, 具有非常好的操作性特点, 通过控制惩罚的类型、方式、强度等, 并引入其他变量, 如互惠、利他、公平和信任等, 揭示二者之间的关系, 观察惩罚对合作的影响作用。公共物品困境中对惩罚的操纵表明: 惩罚搭便车可以维持群体内的合作行为。早期研究表明, 惩罚机制有助于建立社会合作规范, 有代价惩罚、利他惩罚和第三方惩罚等惩罚形式为间接互惠提供了合理的解释。随着实证研究的进一步深入, 惩罚表现出新的形式, 如自私惩罚和反社会惩罚, 被惩罚者的威胁报复将使合作者不再继续惩罚搭便车行为, 导致合作行为的减少直至消失。通过正当性理论、道德情感理论和文化基因协同进化理论表明惩罚的存在具有一定的必要性和价值, 惩罚的作用存在框架效应。本文介绍一些中介变量, 如利他、信任、声誉、文化、期望等, 可以对惩罚作用的积极面和破坏性予以解释, 也为合作的心理机制研究提供新的视角。

2 惩罚的表现形式

Carpenter, Matthews 和 Ong'ong'a(2004)指出惩罚包括惩罚者 (punisher) 和被惩罚者, 根据参与人的捐献数额和对待惩罚的态度, 将惩罚者分为五种基本类型 (punisher types): (1) 搭便车并惩罚合作, 即原则性的搭便车 (principled free riders); (2) 搭便车, 不惩罚; (3) 搭便车并惩罚搭便车, 即伪善的搭便车 (hypocritical free riders); (4) 合作, 不惩罚; (5) 合作并惩罚搭便车, 即原则性的合作 (principled cooperators)。公共物品博弈研究中, 惩罚作为自变量, 具有多样化的形式和操纵方案,

本文以为实验者对惩罚变量的操纵和控制, 表现形式可以归纳为以下四类: (1) 根据由谁来实施惩罚, 分为第一方惩罚、第二方惩罚和第三方惩罚; (2) 根据惩罚的成本, 可分为有代价惩罚和无代价惩罚; (3) 根据惩罚的结果, 可分为利他惩罚、自私惩罚和反社会惩罚; (4) 根据惩罚的内容, 分为金钱惩罚和社会惩罚。

2.1 第一方、第二方和第三方惩罚

公共物品博弈涉及到 N 个参与人, 如果在捐献时, 一名参与人搭便车, 损害另一参与人的利益, 那么搭便车的参与人称之为“第一方”, 合作的参与人称之为“第二方”, 而知晓这个捐献过程, 没有卷入的其他人被称之为“第三方”。第一方惩罚指个人内在的自我良心惩罚, 当个人违背群体规范时产生的不舒服感, 如内疚、难过、羞愧和尴尬。它是建立在人性本善的假设上, 一般来说, 个人在偏离社会规范时感受到良知的呼唤, 这种良知的惩罚会促进个体做出补偿性行为 (Masclét & Villeval, 2008)。

第二方惩罚 (second-party punishment) 指被利用或被欺骗的第二方对给予其损失的第一方的惩罚, 有时也称之为同伴惩罚 (peer-punishment)。代表性的是最后通牒博弈, 一名提议者向一名响应者提出一种资源分配的方案, 如果响应者同意这一方案, 就按照这种方案进行资源分配; 如果不同意, 则两人什么都得不到。响应者用“不同意”来表达对分配不满意的惩罚, 这种惩罚为第二方惩罚。按照理性人假设, 只要提议者将少量资源分配给响应者, 响应者就应该同意, 因为这要比什么都得不到的好。但实验表明提议者倾向于对半分分配, 而响应者少于 30% 时拒绝, 多于 30% 时接受。

第三方惩罚 (third-party punishment) 指旁观者实施的对搭便车的惩罚 (Cameron, Payne, & Knobe, 2010; King & Wheelock, 2007)。显而易见, 国家司法系统和监督系统充当了第三方惩罚的角色, 生活中“路见不平, 拔刀相助”的侠义行为也在此列。对于正义和公平的追求是人类孜孜以求的奋斗目标, 当第三方目睹非正义行为时, 产生一种维护正义的冲动, 这种维护正义的冲动经常被文化所鼓励, 如在《水浒》中称之为“替天行道”。第三方惩罚的存在说明人们维护社会规范的强烈意愿, 也许是基于在目睹他人不合作行为时, 对被欺骗

和利用一方的共情能力,体验到愤怒、无助等消极情绪。de Kwaadsteniet, Rijkhoff 和 van Dijk(2013)的研究表明,大约有三分之二的第三方惩罚搭便车行为。有趣的现象是,第三方惩罚实施的力度往往小于第二方惩罚,当被惩罚无法避免时,被惩罚者倾向于选择第三方惩罚(Fehr & List, 2004)。

2.2 无代价惩罚和有代价惩罚

现实生活中实施惩罚需要付出一定的成本,比如时间、精力或者金钱,人们根据成本的高低来决定惩罚与否,当成本过高时放弃惩罚,不计成本的惩罚可能是被逼入绝境或者是有偏执性格倾向。在公共物品博弈中,在明确提出“有代价惩罚”的概念之前,实验操纵的惩罚多数是无代价的惩罚方式(free punishment),即惩罚者不需要付出个人拥有,就可以减少搭便车的收益。高捐献者自主决定是否对搭便车实施惩罚,以及惩罚力度,高捐献者不需要付出任何代价。Gintis 和 Fehr (2012)比较设置惩罚之前和之后的捐献,发现在设置惩罚之前,约十次左右捐献降到一个较低的水平,并且上次比平均捐献水平高的参与者,这次较少捐献;在设置惩罚之后,平均捐献水平保持在个人拥有的一半代币以上,高捐献情境下,几乎没有人继续使用惩罚。

考虑到成本付出,掀起了惩罚研究的新高潮,对惩罚的操纵出现变化,公共物品困境中有代价惩罚成为主要的惩罚方式(Carpenter & Matthews, 2009; Fehr, Hart, & Zehnder, 2011; Dela Rama, 2012; Ostrom, 2012; von Rueden & Gurven, 2012)。有代价惩罚(costly punishment)是指参与者付出个人账户的一部分或全部以惩罚其他参与者,例如个人支付 n 个代币,就可以减少其他人的收益。Ostrom (2012)的研究表明有代价的惩罚对维持合作具有明显的作用,人们宁愿付出代价惩罚搭便车,它具有以下三个操作性特征:(1)是一种个人自愿的行为模式,不受权威人士的干涉,不同于社会的法律法规;(2)个人需付出一定成本以购买惩罚权,是一种有代价的惩罚形式;(3)其结果是减少受惩罚对象的所得利益。

2.3 利他惩罚、自私惩罚和反社会惩罚

2.3.1 利他惩罚

利他惩罚(altruism punishment)是指个人表现出合作行为,并且惩罚那些搭便车者,即使彼此

只有一次合作机会,个人宁愿付出一定的代价也要惩罚搭便车(Rebers & Koopmans, 2012)。无论是单次还是多次公共物品博弈,人们表现出利他惩罚的倾向。惩罚的目的是警示和威慑,要求搭便车者在此后表现出合作行为,惩罚的结果具有一定程度上的利他性。当法律和法规不能鼓励合作行为时,利他惩罚解释了为什么在遗传上没有关联的陌生人之间能保持对社会有益的高水平合作,利他惩罚可能是合作进化中决定性的力量。不过,也有不同的观点,Taggar 和 Neubert (2008)认为原始社会人们生活在一个彼此熟悉的小范围内,在群体内实施匿名惩罚几乎是不可能的,也不可能学会适应陌生人之间互相惩罚的情境。并且,惩罚付出的代价和被利用付出的代价是不对等的,惩罚的成本会降低惩罚者的适应性(fitness),与惩罚的合作者相比,不惩罚的合作者成为一种“次级搭便车”(second-order free riders),这使得利他惩罚成为一个稳定的进化策略似乎有些困难(Shutters, 2009)。

2.3.2 自私惩罚

自私惩罚(selfish punishment)指搭便车对搭便车的惩罚。在缺乏透明的监督机制时,公共物品博弈中的少数自私者表现为“贼喊捉贼”,以混淆视听和判断,逃避被惩罚。人们喜欢惩罚搭便车,同样也喜欢欺骗和隐瞒,这种伪善的有代价惩罚使得惩罚行为变成是一种策略性的选择。因为搭便车减少群体中每个人的适应性,包括其他搭便车,所以搭便车有很强烈的超过互惠的动机除掉其他搭便车,“贼喊捉贼”暗示着没有人比骗子更能发现骗子,因此在骗子之间也存在相互竞争。骗子具有专门的优势来发现和识别其他骗子,他们可能对行骗的策略更为熟悉或者具有丰富的斗争经验。一个极端的利他主义者可能是最富有掩饰性的自私行为,是有能力的欺骗者,也许能排斥其他的骗子而成为骗子中的极品(Kiyonari & Barclay, 2008)。持这种观点的研究假定惩罚是自私的策略在起作用,高明的搭便车为了自身利益的最大化,惩罚搭便车,极力维持和保护“合作羊群”。自私惩罚比利他惩罚能更好的解释利他的发展,利他能得到进化是基于欺骗者之间的竞争,因此利他与惩罚是无关的,惩罚只可能是自私的举措。

2.3.3 反社会惩罚

反社会惩罚(antisocial punishment)是指个人

不惩罚搭便车，而惩罚表现出合作的人(Laurent, Masclet, & Noussair, 2007)。亲社会倾向的人表现出更多的合作行为(Yamagishi et al., 2013; 王沛, 陈莉, 2011)，而反社会倾向的人更多搭便车，变本加厉的是还可能对亲社会行为表达出不屑和反抗。比较了世界范围内不同文化下的被试群体，发现反社会惩罚是一个跨社会变量(cross-societal variation)，在一些群体中，反社会惩罚的力量如此强大，以至于能抵消惩罚对合作的增强效果。同时，也发现国家法制水平和公民合作规范可以很好地预测反社会惩罚，并且只有当惩罚是社会合作规范的补充时，惩罚才有利于社会(Henrich et al., 2005)。

反社会惩罚可能基于“以牙还牙”的复仇心理。复仇被看成是一种不计代价或风险，施加惩罚于那些曾经使之痛苦的人身上。这意味着不仅仅是合作者希望惩罚搭便车，同时搭便车也有强烈的对惩罚给予反击的念头。人们不仅有复仇的动机，而且使用反惩罚的策略阻碍以后的惩罚(Milinski & Rockenbach, 2008)。并且，在目前的实验设计中，惩罚者付出的代价和被惩罚者的损失是不对等的，惩罚者占优势，被惩罚者损失得更多，对给予自己惩罚的报复就成为反社会惩罚的动机。有近四分之一的被试会对惩罚给予报复，并且反惩罚对合作极具破坏性，被惩罚者关于报复的威胁削弱合作者惩罚搭便车的意愿，最终导致合作的崩溃(Nikiforakis, Normann, & Wallace, 2010)。

2.4 金钱惩罚和社会惩罚

惩罚的内容可以是物质上的，也可以是精神上的。俗话说：“天下熙熙，皆为利来；天下攘攘，皆为名往。”公共物品博弈中，通过减少被惩罚者的代币收益来实现的惩罚，称之为金钱惩罚，因其便于量化，是目前实验研究最常用的(Mulder, van Dijk, De Cremer, & Wilke, 2006; Mulder, Verboon, & De Cremer, 2009; Baumgartner, Knoch, Hotz, Eisenegger, & Fehr, 2011; Yamagishi et al., 2013)。此外，还有一种惩罚形式为社会惩罚，比如批评、谩骂、孤立、污名化等，社会形式(也有称言语形式)惩罚研究相对较少，但是仍然有一些研究指出批评(disapproval)和社会赞许(social approval)，在单独或者结合社会认同感等条件下，能够提高个体在社会困境中的合作水平(Mulder

et al., 2009; 韦倩, 2009)。社会赞许意味着个体被他人接受和喜爱，这能够带来满意、快乐、自豪感等积极情绪；反之，面对他人的批评或者羞辱，容易引起个体尴尬、羞愧、沮丧、难过等消极情绪。社会形式的惩罚，如流言蜚语(gossip)提供了一种以社会赞许或者社会压力形式的合作保障(Axelrod, 1980; Guala, 2012)。

在金钱惩罚、社会惩罚和两种惩罚混合使用的三种条件下，被试更多选择社会惩罚，并且相对与金钱惩罚，当社会批评撤除后，个体在惩罚取消阶段的合作和信任水平下降的程度较大(Guala, 2012; 刘谓, 马剑虹, 朱玥, 2010)。这表明社会惩罚的作用似乎大于金钱惩罚，并且社会惩罚的成本相对较低。当金钱被认为是身外之物时，金钱惩罚对自尊的破坏力几乎可以忽略，社会惩罚可能对个体的自尊有着较大的破坏力(王沛, 陈莉, 2011)。经济所得和社会接纳被认为是人们行为的基本动机，在保证人们身心健康的同时，如何在两者之间进行选择或是结合以发挥惩罚的效力成为研究进一步深入的方向。

3 与惩罚有关的研究

3.1 惩罚与合作

公共物品困境中惩罚搭便车可以维持合作行为，得到了很多研究的支持(Rosecrance, 1981; Fehr & Gächter, 2000; Shinada & Yamagishi, 2007)。实验研究表现为三种模式：(1)简单比较实验组和控制组，一组给予惩罚，一组没有惩罚。(2)“移开惩罚”范式，即实验组包括两个阶段，增加一个移开制裁的阶段。(3)有代价惩罚机制。研究表明，假如让被试在无惩罚和有惩罚情境中选择，在实验开始阶段，多数被试选择没有惩罚的系统，随着搭便车行为的增多，出于对被利用的厌恶，无惩罚机制下的合作逐渐减少到零。接着，面对重新选择，几乎所有被试都选择有惩罚的情境，在惩罚机制下，合作保持在一个均衡的水平(Mulder et al., 2006)。惩罚维持合作的原因在于惩罚是一种强制力量，它对搭便车的惩罚，使得搭便车行为成为不具吸引力的选择，提供人们对其他人合作的信任和期望。这些研究说明：(1)社会困境中惩罚的存在是必要的；(2)个人是有限理性的；(3)个人的惩罚意愿不一致。惩罚搭便车维持合作被证实后，研究普遍关注惩罚是如何影响合

作的,其他一些相关实证研究为惩罚与合作的关系提供了翔实的内容。

3.2 惩罚与利他

利他惩罚及其认知神经基础的研究成为这几年的热点。即使惩罚是必须付出成本和代价的,大多数人仍然用惩罚来表达对搭便车的不满,不实施惩罚比惩罚有更好的收益,有代价惩罚的结果似乎是一种利他,利他惩罚这个概念也应运而生(Fehr, Fischbacher, & Gächter, 2002)。利他惩罚主要涉及以背外侧前额叶皮层为主的基于规范的系统和包括纹状体、内侧前额叶皮层以及脑岛等脑区在内的情绪驱动系统。此外,五羟色胺、多巴胺等神经递质可能是利他惩罚发生的神经生化基础。“认知控制”与“情绪满足”被用来阐述利他惩罚的产生机制(Shutters, 2009; Baumard, 2012)。

当被试可以真正的或者只是象征性的惩罚搭便车时(真正的惩罚可以减少搭便车的盈利,象征性的惩罚不会损害搭便车的盈利),扫描被试的脑部,真正的惩罚比象征性的惩罚更强烈地激活尾核与壳核。人类自激励机制的行为是由中脑系统的尾核与壳核来执行的,其激活说明对违背规范的惩罚本身能让人感到满意(Bernhard, Fischbacher, & Fehr, 2006; Adams & Mullen, 2012; Guala, 2012; Skatova & Ferguson, 2013)。不过,反社会惩罚和自私惩罚的存在,说明有代价惩罚的初衷可能不是简单的利他,可能还有其他原因,仍需进一步探讨。在探寻人类为什么合作时,社会生态学认为利他惩罚维护了社会规范,强有力地展示了公平规范和亲社会行为的价值(Rebers & Koopmans, 2012)。

3.3 惩罚与信任

信任(trust)是指对他人的行为和动机的积极预期,这种预期往往与个人屡次参与活动的经历有关,互惠是这个过程的核心。信任可以细分为人际信任和制度信任,其中人际信任与个体的亲社会价值取向有关,是内在的信任;制度信任与社会法规有关,是外在的信任(马剑虹,徐美玲,2011)。一般来说,当群体内的人际信任水平较低时,人们选择设置惩罚(Yamagishi, Kikuchi, & Kosugi, 1999)。不守规矩就得受惩罚,惩罚设置提供了一种制度信任的保障,惩罚的威慑作用,使得人们相信惩罚可以减少搭便车,捐献水平保持在一个较高水平。同时,由于惩罚具有强制力量,

强调人们必须合作,而不是自愿选择是否合作,它剥夺了人们自由做出选择的权利,破坏了合作愿望,容易导致心理阻抗,可能破坏成员之间的人际信任水平,从而拒绝合作。

Tenbrunsel 和 Messick (1999)的研究表明,惩罚的强度也会影响合作的水平,比如无惩罚、轻微惩罚和重度惩罚三种条件下,与无惩罚相比较,轻微惩罚下的合作水平不升反而降低。这表明所谓轻微惩罚反而适得其反,轻微的惩罚设置降低了人际信任的水平,又无力构建制度信任的堤坝。王沛和陈莉(2011)发现惩罚削弱人际信任水平,经历过惩罚的被试的人际信任水平低于没有经历过惩罚的被试,经过中介效应检验发现,人际信任在惩罚影响亲社会价值取向被试的合作程度中起中介作用。惩罚通过亲社会型博弈者的人际信任水平对合作程度有间接负效应,即惩罚强度越强,人际信任水平越低,合作水平也越低。存在惩罚可以保持合作,基于惩罚表明制度信任的可信度,影响制度信任的水平,信任是惩罚与合作的中介变量之一(Mulder et al., 2006)。

3.4 惩罚与声誉

声誉(reputation)是一种积极的社会评价,对个体而言,意味着正义、公平与合作的名声,声誉本身就是一种收益。惩罚搭便车的举措,虽然暂时损失了物质利益,但是收到了好的声誉。与社会性惩罚相类比,具有良好声誉有助于利他惩罚的推广(Baumard, 2010)。这表明惩罚的最终效果不是直接减少搭便车的物质利益,而是通过这种举措证明:惩罚是正当的,不只是为了一己私利。这样就为实施惩罚营造声誉,声誉系统的建立对个人来说是有好处的,当群体内成员间实施惩罚变成普遍行为时,惩罚者与不惩罚者相比较,在收益处于不利的位置,但惩罚搭便车对群体来说是利好的行为,那么,惩罚者便发出了一种信号:惩罚者是可信赖的,是愿意对群体承担义务的,是关心公平和不想被欺骗的。

如果惩罚是可信赖的和公平的信号,那么惩罚者虽然损失了一些物质利益,但可以从那些除自私者以外的独立者那里得到好的声誉,人们希望与那些表明自己是不能容忍欺骗或偏袒行为的人建立合作关系的,因此,惩罚者比非惩罚者能得到更具有合作意识的伙伴。实验结果表明,在利他惩罚结束后,请被试对实施利他惩罚的惩罚

者进行声誉评定,惩罚者被认为是更可靠、关注群体利益和值得尊重的(Bernhard et al., 2006)。在某种程度上,社会规范的沿袭是以维持声誉为基础得以保留的,丢失声誉可以被看作是违反社会准则,声誉机制使得类似于合作之类的社会规范得以在集体中体现和保持下去。

3.5 惩罚与文化

惩罚与文化密切的联系体现在两种研究取向上,一种关注在不同社会文化下,惩罚的有效性和人们对惩罚的认同;另一种是惩罚如何体现了人类文化的变迁和演化。例如,人类从自由文化发展到制度文化的过程,制度文化如何在人类社会中立足和完善起来的呢?公共物品博弈提供了一种情境,在第三方惩罚制度尚未正式建立时,有代价惩罚搭便车的强烈意愿,为制度文化的建立提供了一种解释。假如在一个有惩罚的社会环境里,违反社会规范的人被惩罚,遵守规范的人其利益得到保护,在文化中首先显现出惩罚的价值来,个人为避免惩罚可能变得更具有合作意识,利他惩罚是建立人类合作规范、文化制度和亲社会行为可行的解释。正式制度对不合作行为的惩罚效果是稳定的,正式制度一旦确定,合作不再是自主自愿的行为,而是一种被强制的表现,阻止不同意见或者是少数派意见的表达(Olsson-Yaouzis, 2012)。韦倩(2009)指出社会中存在能够增强人类惩罚能力的三种社会机制,如规范的内化、缔结同盟与第三方接入,这三种社会机制对利他惩罚行为演化的影响,表明了它们在维持人类社会合作中的重要作用。反社会惩罚的概念也是在跨文化研究中提出来的,从不同的国家收集数据,包括不同经济发展水平和制度建设的国家,发现反社会惩罚与社会的政治体制、人权、公民自由、腐败等制度建设呈负相关,民主水平越高,反社会惩罚越少。

4 惩罚的作用

设置惩罚有三个理由:第一,惩罚比劝说更有效,惩罚是昂贵的,劝说是廉价的,比如“禁止搭便车”与“搭便车罚款”的警示,罚款更能表达对搭便车的不满,所以说惩罚比规范更具直观性,惩罚是基于道德上的义愤或者内心直觉,不是科学的逻辑的分析,它可能是以偏概全的,表现出启发式的思维习惯。第二,对于不便于量化或难

以具体化的社会性规范,惩罚操纵可以有一些弹性空间。惩罚直观性、道德义愤和认知偏差这些特点,使得实施惩罚具有情境性和灵活性,如“法不责众”就有这层涵义。第三,惩罚者能从惩罚违规行为中有所获益,惩罚具有道德义愤感,被看成是合法的工具,一旦使用就可能获得金钱补偿、声誉补偿或者是兼而有之(de Kwaadsteniet et al., 2013)。

与无惩罚相比较,惩罚条件下参与人更多地选择合作。这表明通过对搭便车的惩罚,加强了不确定决策中个体的合作水平。这提供一种期待——倡导合作,与搭便车相比较,合作是更有吸引力的选择。比如“路边随意停车罚款二百元”目的是为了维护交通安全让司机规范停车,不随意停车,而不是为了多收罚款。人们选择惩罚来管理和控制行为,惩罚是为了引导和增强规范行为,形成社会规范,本文试图从以下的四个方面来介绍惩罚的作用:

4.1 正当性理论

哲学上通过效用理论(utility theory)和报应理论(retribution theory)论证惩罚存在的必要性,指出实施惩罚是正当的(justness)。效用理论认为惩罚通过改变个人“成本—收益”的计算权衡来威慑潜在的搭便车,对于自私自利者而言,从公共物品中获得的收益超过提供的成本时,为了确保收益就会提供适当的成本。理性人追求个人效用的最大化和群体利益的均衡,从而保障了双方的权益,惩罚能使合作成为一个占优的选择。边沁说:“所有惩罚都是损害,所有惩罚本身都是恶,根据功利原理,如果允许惩罚,那只能是因为惩罚可以排除某种更大的恶。”惩罚不仅具有阻止、改造和威慑搭便车行为的能力,而且对搭便车的惩罚具有谴责和示范效应,惩罚的威慑能阻止搭便车行为的重复出现,所以惩罚是正当的。

报应理论则认为不是由于惩罚能取得什么好的结果,而是对被惩罚者而言,惩罚是应得的一种报应,“善有善报,恶有恶报”和“以德报德,以怨抱怨”的因果报应似乎是理所当然的,惩罚是面对破坏规范理所当然的反应,不关注惩罚任何功利的后果。康德说:“惩罚在任何情况下,必须只是由于一个人已经犯了一种罪行才加刑于他。”根据恶行的邪恶程度(包括根据搭便车造成伤害的大小、搭便车的意图等)来给予一定量度的惩罚,

这在道德上是恰当地,是为了维护人类的正义,所以惩罚是必须的、合理的(王立峰,2012)。关于合作进化的观点中,以囚徒困境为实验材料,计算机仿真的研究表明:“一报还一报”(tit for tat)是行为博弈中较好的合作策略,说明惩罚的正当性理论具有一定的解释作用。(Axelrod, 1980)。

4.2 道德情感理论

道德情感理论(moral sentiment model)认为惩罚是人们面对不公平现象时一种自发的、自然的情绪反应,例如愤怒、报复等(Nikiforakis et al., 2010)。亚当·斯密提出人性中具有一种同情的能力,“将心比心”被认为是道德情感中的黄金定律,即一些事情我不能对别人做,因而推测别人也不会对我这样做,即人类先天具有亲社会性,使得人们在面临利益冲突的情形下考虑道德要求,如诚实、同情、可信赖、体谅和忠诚,这些道德标准可以缓和搭便车的冲动。例如,认为低捐献者是自私的人,他们在被惩罚后增加捐赠是为了避免以后的惩罚。但很多低捐献者对象征性的惩罚也做出增加捐献的反应,这表明他们可能不是自私的,而是为避免被利用或者因违反捐献规范感到羞愧,所以增加捐献。

情感理论包括情感反应极端形态模型(the extreme form of emotional reaction model)和不平等厌恶模型(inequality aversion model)。情感反应极端形态模型认为惩罚是愤怒的结果,或者是类似于愤怒的某种情绪的结果,人们认为在捐献条件下互惠或者利他是应当服从的规范,自私行为会招致惩罚(Baumard, 2010, 2011)。对搭便车的惩罚是因为对公平规范的违背和感觉到不满产生的愤怒情绪,愤怒是一种合作者对搭便车的消极情绪,是惩罚的最近的机制。不平等厌恶模型认为惩罚者有其他不同的选择,即惩罚效用的最大化。人们厌恶不公平,尤其是当局面对自己不利时,即使在惩罚的过程中自己的得益减少了,但是通过减少不公平,仍可以维持自身的效用,这个模型认为若高捐献者十分反对不平等,即使惩罚的成本可能超过了惩罚者自己的所有,高捐献者也将惩罚低捐献者。

4.3 文化基因协同进化理论

惩罚是通过社会学习获得的一种文化适应行为,符合文化基因协同进化的规律(Fehr et al., 2002; Ostrom, 2012)。协同进化(culture-gene

co-evolutionary model)包括社会学习理论和昂贵的信号传递理论(costly signaling model)。社会学习过程中的遵奉传播(conformist transmission)能建立惩罚,人类不只是简单地模仿他们的父母,也不是随意模仿他人,而是使用社会学习的两个基本规范:“复制成功”(copy the successful)和“复制大多数”(copy the majority),即复制最成功个人和人群中最频繁行为。一旦惩罚与合作的关系建立,它能通过进化机制——团体选择的方式被广泛传播,例如,一个稳定合作均衡环境的团体更具有竞争力,处于占优地位,那么,占优就会使惩罚与合作的行为扩散到个人层面,导致更多的复制与模仿,引发基因上亲社会行为的增生、扩散,文化基因协同进化减少个人适应环境的成本,是一个低成本高效率的方法,通过社会学习高效的获得适应性行为。

生物学中“昂贵的信号传递理论”通过考察行为传递的信息解释了明显的非适应性的行为或突变。例如,雄性动物总是发展出一些看起来像是缺陷的生理特征,比如,雄性孔雀有着无与伦比的美丽羽毛,在求偶时雄孔雀通过开屏来展示魅力,美丽的羽毛并没有其他适应功能。而缺陷传达给雌性的信息却是,该雄性如此强壮以至于它能承受这种负担,它是有吸引力的。据此,惩罚虽然是件耗时费力的举措,但是它传递一种信号,表明一种态度“我不是好利用的”。所以说,惩罚具有信号传递的功能,参与者似乎能意识到他们选择的信号能够传递信息,并且接受者从信号中做出推理。个人采用较多捐献提供了“合作”的信号,反社会惩罚提供了“反社会报复”的信号。惩罚作为一种可识别信号,它降低人际之间信息交易的成本。那么,这种行为很快就会被人们复制、模仿、传播和稳定下来。所以文化基因协同进化理论也为惩罚提供了一种解释。

4.4 “伦理-交易”框架效应

框架效应(framing effect)是对相同决策任务的不同描述方式会影响决策者对备择选项的表征,从而影响其决策偏好的现象,其经典的例子是亚洲疾病问题。经典框架效应中,“框架”主要是对金额做了“得”或“失”的不同表征,因此,经典框架效应也可以称之为“得失框架效应”(马文娟等, 2012)。公共物品困境中的决策偏好也表现出框架效应,对框架的敏感性影响决策,一种是给框架

(give frame), 参与人打算向公共账户捐助多少; 另一种是拿框架(take frame), 参与人从公共账户中能获益多少。

框架形式表现出新的内涵, 这种框架效应不同于传统的“给—拿”概念, 而是提供不同的参照系统作为决策的偏好。当设置惩罚时, 公共物品困境中, 决策偏好不再仅仅表现“给—拿”框架效应, 而是表现出“伦理-交易”框架效应。在人们的头脑里, 伦理道德框架(ethical decision frame)和市场交易框架(business frame)共同存在, 两种框架下对损失和收益的认识是不同的。在伦理道德框架中, 符合伦理规范是个人的优先选择, 但是在市场交易框架中, 成本收益是优先的选择。

惩罚设置提供一种可识别的社会线索, 启发式的思维方式会对惩罚诸如此类的词语较为敏感, 与伦理决策框相比较, 存在的惩罚减少了人们对伦理道德框架唤起的可能性, 增加了市场交易框架被唤起的可能性(Cubitt, Drouvelis, & Gächter, 2011)。“取消惩罚”的研究范式表明, 在设置惩罚之后再移开惩罚, 与刚开始就没有设置惩罚的条件相比较, 前者的合作水平明显下降, 这说明惩罚已经破坏了人际之间的内在信任(Fehr & Schmidt, 1999; Mulder, van Dijk, Wilke, & De Cremer, 2005)。这也验证了个体对惩罚的理解存在框架效应, 惩罚的设置使得合作决策成为一种策略性的成本收益考虑, 而不是基于伦理道德的潜在要求。也就是说, 惩罚改变了人们的合作动机, 从道德的、伦理的动机变成计算的、交易的动机。假如参与人决策时知道存在某种惩罚, 就会认为决策是一种市场交易, 精于得失的算计而忽略了其中的道德内容。因此, 惩罚的作用体现在: (1)当惩罚的成本超过搭便车的收益时, 人们将合作以避免惩罚; (2)当惩罚的成本低于搭便车的收益时, 人们将支付罚款并使个人利益最大化; (3)惩罚缺失时, 人们将在与伦理有关的情境下做决策, 相对而言完全自私的决策很少发生(陈欣, 叶浩生, 2007)。

5 小结与展望

5.1 惩罚作用的两面性

惩罚表现出积极的作用, 似乎又具有潜在的破坏性。早在操作条件反射中, Skinner 就提出惩罚的操作性概念, 即在不良行为产生后呈现一个

厌恶刺激或取消一个愉快刺激, 使其行为发生概率减少。他指出惩罚能够引起有机体生活或者行为反应的变化。但是 Skinner 反对使用惩罚, 认为在多数情况下惩罚是不好的, 一是它并不一定能削弱不良行为, 不能很好的实现对不良行为的控制。受到惩罚的不良行为, 很可能在惩罚消失后重新出现。二是惩罚指出不应该做什么, 却无法告知应该做什么, 而且无形之中承认了让他人遭受痛苦行为的合理性。使用惩罚往往引发负面情绪, 容易引起其他不良的反应, 例如攻击敌意行为。他相信正强化的效果, 希望能找到代替惩罚的方法, 比如: 忽视不良行为或者是改变不良行为出现的外部环境。

由于操作的直观性、操作性和有效性, 使其在人类生活中一直扮演着重要的角色。在利他惩罚之外, 发现自私惩罚和反社会惩罚等新现象, 表明惩罚确实会引发一些敌意行为, 但惩罚具有维持合作的积极作用似乎更引人注目。那么人类真的能通过惩罚来管理自身吗? 惩罚的最终目的是什么? 一种观点认为惩罚的目的是为了没有惩罚, 当人们都遵守一些共同的价值信念行事时, 惩罚就变得没有必要了, 现在的惩罚是为了以后不再有惩罚。反对的声音则认为, 奖励和惩罚就像是一枚硬币的两面, 奖惩的思维方式是对多元化的一种扼杀, 一致的、标准的行为模式只会导致人类生活的单调和无趣。

值得一提的是, 由于实验设计具有将复杂问题简单化倾向, 这使得被试很容易推测实验设计的目的, 从而洞察实验的主题, 影响结果解释的范围。因此, 实验者对惩罚的操纵更类似于方法论上的设置。惩罚作为一种设置, 具有情境性, 可以是强行的规定, 也可以是一种自愿的选择, 还可能是自发自觉的行为, 或是群体协商的结果, 而不得是一种符合现实生活的社会偏好。虽然惩罚对维持群体内的合作是必要的, 但是脱离情境对惩罚的作用进行评价仍是困难的。

5.2 惩罚研究的前景

公共物品困境中惩罚的操纵颇具吸引力, 社会学研究关注惩罚的效果与文化信念、社会制度、法律法规的联系, 经济学关注人的理性和利益得失之间的博弈, 生物学则强调惩罚也许是人类合作进化的动力来源以及其所具有的生理基础, 教育学则关注惩罚对学生造成的身心伤害, 心理学

更多是对惩罚的认知过程与神经基础感兴趣。公共物品困境的信任、共情和偏好,作为惩罚与合作关系研究的中介变量,这些问题的研究带有浓厚的心理学特色,在上述的惩罚与合作、利他、信任、声誉和文化等变量之间的关系研究之外,其研究前景还表现在以下三个方面:

一是对惩罚作用两面性的继续探讨。不同形式和强度的惩罚在哪些情况下发挥积极作用,哪些情况下表现出破坏性,或者是引入中介变量,对哪些变量表现为积极作用,哪些变量表现为消极作用。

二是惩罚研究中参与人体验到的诸多情感和情绪。尤其是一些具体的感受(specific emotions),例如满意、愤怒、敌意、报复、内疚等。在愤怒敌意的情绪之下人们花很大的代价实施对他人的惩罚,在内疚后悔的情绪之下,人们会自我批评或惩罚。情感是个体对人对己或事件的一种直接、快速的反应,与深思熟虑的理性相比较,情感是凭直觉获知的,研究注意到情感具有潜在的强大力量,具有重要的研究价值。

三是惩罚研究中涉及到的对他人行为的推断,这种推断存在期望差异,所谓期望差异(expectation discrepancy)是个体与个体之间因为立场、角色差异而存在一些容易出现又很难调和的知觉差异,这种知觉差异容易导致人际冲突(谢晓非,胡天翊,林靖,路西,2013)。公共物品困境中,因参与人角色、对信息的知觉、知识水平、心理权利和信任等因素之间的差异,导致沟通障碍和合作困难,造成惩罚的使用或滥用,减少期望差异或许能避免滥用惩罚,这值得进一步的关注。

参考文献

- 陈欣,叶浩生.(2007). 两难中合作行为研究的回顾和展望. *心理科学进展*, 15(5), 743-748.
- 刘谓,马剑虹,朱玥.(2010). 从归因视角探讨公共物品两难中惩罚系统对合作的影响. *应用心理学*, 16(4), 332-340.
- 马剑虹,徐美玲.(2011). 信任及其对公共财物两难博弈中合作行为影响. *应用心理学*, 17(1), 10-17.
- 马文娟,索涛,李亚丹,罗笠铎,冯廷勇,李红.(2012). 得失框架效应的分离——来自收益与损失型跨期选择的研究. *心理学报*, 44(8), 1038-1046.
- 王立峰.(2012). 论惩罚的政治正当性. *中国刑事法杂志*, (5), 3-12.
- 王沛,陈莉.(2011). 惩罚和社会价值取向对公共物品两难中人际信任与合作行为的影响. *心理学报*, 43(1), 52-64.
- 韦倩.(2009). 增强惩罚能力的若干社会机制与群体合作秩序的维持. *经济研究*, (10), 133-143.
- 谢晓非,胡天翊,林靖,路西.(2013). 期望差异:危机中的风险沟通障碍. *心理科学进展*, 21(5), 761-774.
- Adams, G. S., & Mullen, E. (2012). The social and psychological costs of punishing. *Behavioral and Brain Sciences*, 35(1), 15-16.
- Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *The Journal of Conflict Resolution*, 24(1), 3-25.
- Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind & Society*, 9(2), 171-192.
- Baumard, N. (2011). Punishment is not a group adaptation. *Mind & Society*, 10(1), 1-26.
- Baumard, N. (2012). The restorative logic of punishment: Another argument in favor of weak selection. *Behavioral and Brain Sciences*, 35(1), 17-18.
- Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C., & Fehr, E. (2011). Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nature Neuroscience*, 14(11), 1468-1474.
- Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, 442(7105), 912-915.
- Cameron, C. D., Payne, B. K., & Knobe, J. (2010). Do theories of implicit race bias change moral judgments? *Social Justice Research*, 23(4), 272-289.
- Carpenter, J., & Matthews, P. H. (2009). What norms trigger punishment? *Experimental Economics*, 12(3), 272-288.
- Carpenter, J. P., Matthews, P. H., & Ong'ong'a, O. (2004). Why punish? Social reciprocity and the enforcement of prosocial norms. *Journal of Evolutionary Economics*, 14(4), 407-429.
- Cubitt, R. P., Drouvelis, M., & Gächter, S. (2011). Framing and free riding: Emotional responses and punishment in social dilemma games. *Experimental Economics*, 14(2), 254-272.
- de Kwaadsteniet, E. W., Rijkhoff, S. A. M., & van Dijk, E. (2013). Equality as a benchmark for third-party punishment and reward: The moderating role of uncertainty in social dilemmas. *Organizational Behavior and Human Decision Processes*, 120(2), 251-259.
- de Quervain, D. J. F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, 305(5688), 1254-1258.
- Dela Rama, M. (2012). Corporate governance and corruption: Ethical dilemmas of Asian business groups. *Journal of Business Ethics*, 109(4), 501-519.

- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4), 980–994.
- Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature: An Interdisciplinary Biosocial Perspective*, 13(1), 1–25.
- Fehr, E., Hart, O., & Zehnder, C. (2011). Contracts as reference points-experimental evidence. *The American Economic Review*, 101(2), 493–525.
- Fehr, E., & List, J. A. (2004). The hidden costs and returns of incentivetrust and trustworthiness among CEOs. *Journal of the European Economic Association*, 2(5), 743–771.
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817–868.
- Gintis, H., & Fehr, E. (2012). The social structure of cooperation and punishment. *Behavioral and Brain Sciences*, 35(1), 28–29.
- Greenberger, D. B., Miceli, M. P., & Cohen, D. J. (1987). Oppositionists and group norms: The reciprocal influence of whistle-blowers and co-workers. *Journal of Business Ethics*, 6(7), 527–542.
- Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, 35(1), 1–15.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ... Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795–815.
- King, R. D., & Wheelock, D. (2007). Group threat and social control: Race, perceptions of minorities and the desire to punish. *Social Forces*, 85(3), 1255–1280.
- Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, 95(4), 826–842.
- Laurent, D., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, 33(1), 145–167.
- Liebrand, W. B. G., Messick, D. M. E., Yamagishi, T., & Hayashi, N. (1996). *Selective play: Social embeddedness of social dilemmas*. Heidelberg and New York: Springer.
- Masclet, D., & Villeval, M. C. (2008). Punishment, inequality, and welfare: A public good experiment. *Social Choice and Welfare*, 31(3), 475–502.
- Milinski, M., & Rockenbach, B. (2008). Human behaviour: Punisher pays. *Nature*, 452(7185), 297–298.
- Mulder, L. B., van Dijk, E., De Cremer, D., & Wilke, H. A. M. (2006). Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas. *Journal of Experimental Social Psychology*, 42(2), 147–162.
- Mulder, L. B., van Dijk, E., Wilke, H. A. M., & De Cremer, D. (2005). The effect of feedback on support for a sanctioning system in a social dilemma: The difference between installing and maintaining the sanction. *Journal of Economic Psychology*, 26(3), 443–458.
- Mulder, L. B., Verboon, P., & De Cremer, D. (2009). Sanctions and moral judgments: The moderating effect of sanction severity and trust in authorities. *European Journal of Social Psychology*, 39(2), 255–269.
- Nikiforakis, N., Normann, H. T., & Wallace, B. (2010). Asymmetric enforcement of cooperation in a social dilemma. *Southern Economic Journal*, 76(3), 638–659.
- Olsson-Yaouzis, N. (2012). An evolutionary dynamic of revolutions. *Public Choice*, 151(3–4), 497–515.
- Ostrom, E. (2012). Experiments combining communication with punishment options demonstrate how individuals can overcome social dilemmas. *Behavioral and Brain Sciences*, 35(1), 33–34.
- Rebers, S., & Koopmans, R. (2012). Altruistic punishment and between-group competition. *Human Nature: An Interdisciplinary Biosocial Perspective*, 23(2), 173–190.
- Rosecrance, R. (1981). Reward, punishment, and interdependence. *The Journal of Conflict Resolution*, 25(1), 31–46.
- Rubin, S. S. (1986). Cheating, ethics, and graduate training in professional psychology: Crime and punishment or misjudgment and repair. *Professional Psychology: Research and Practice*, 17(1), 10–14.
- Shinada, M., & Yamagishi, T. (2007). Punishing free riders: Direct and indirect promotion of cooperation. *Evolution and Human Behavior*, 28(5), 330–339.
- Shutters, S. T. (2009). Strong reciprocity, social structure, and the evolution of fair allocations in a simulated ultimatum game. *Computational and Mathematical Organization Theory*, 15(2), 64–77.
- Skatova, A., & Ferguson, E. (2013). Individual differences in behavioural inhibition explain free riding in public good games when punishment is expected but not implemented. *Behavioral and Brain Functions*, 9(1), 3.
- Taggar, S., & Neubert, M. J. (2008). A cognitive (attributions)-emotion model of observer reactions to free-riding poor performers. *Journal of Business and Psychology*, 22(3), 167–177.
- Tenbrunsel, A. E., & Messick, D. M. (1999). Sanctioning systems, decision frames, and cooperation. *Administrative Science Quarterly*, 44, 684–707.

- von Rueden, C. R., & Gurven, M. (2012). When the strong punish: Why net costs of punishment are often negligible. *Behavioral and Brain Sciences*, 35(1), 43–44.
- Yamagishi, T., Kikuchi, M., & Kosugi, M. (1999). Trust, gullibility, and social intelligence. *Asian Journal of Social Psychology*, 2(1), 145–161.
- Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., ... Simunovic, D. (2013). Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes*, 120(2), 260–271.
- Yamagishi, T., & Sato, K. (1986). Motivational bases of the public goods problem. *Journal of Personality and Social Psychology*, 50(1), 67–73.

The Forms and Functions of Punishment in Public-goods Dilemmas

CHEN Xin¹; ZHAO GuoXiang¹; YE HaoSheng²

(¹ College of Education and Science, Institute of Psychology & Behavior, Henan University, Kaifeng 475004, China)

(² Center for Mind and Brain, Guangzhou University, Guangzhou 510006, China)

Abstract: Study on manipulation of punishments in public-goods dilemmas demonstrates that punishment on free-riders can maintain the cooperation behavior among the community. Previous researchers found that the building up of social cooperation norms benefit from the punishment system, in which costly punishment, altruism punishment and third party punishment, etc., provide reasonable explanations to indirect reciprocal. Recent empirical researches illustrate some new forms of punishment, such as selfish punishment and antisocial punishment. However, it was presumed in these researches that revenge from punished free-riders will prevent cooperators to punish them anymore, which would decrease or even dispel cooperation behavior. The necessity of the existence and the value of punishment have been validated by justness theory, moral sentiment model, and culture-gene co-evolutionary model, however, framing effect still exists in the function of punishment. The positive and negative effect of punishment should be further explored by introducing some mediating variables, such as altruism, trust, reputation, culture, and reputation, etc., which could also provide new vision to psychological mechanism of cooperation.

Key words: public-goods dilemmas; free-riders; punishment; cooperation